

Yin Scores and *Yang* Scores: A New Method for Quantitative Diagnostic Evaluation in Traditional Chinese Medicine Research

HELENE M. LANGEVIN, M.D., L.Ac.,¹ GARY J. BADGER, M.S.,² BONNIE K. POVOLNY, M.S., L.Ac.,³
ROBERT T. DAVIS, M.S., L.Ac.,³ ALEXANDER C. JOHNSTON, M.T.C.M.,⁴
KAREN J. SHERMAN, Ph.D., M.P.H.,⁵ JANET R. KAHN, Ph.D., L.M.T.,⁶
and TED J. KAPTCHUK, O.M.D.⁷

ABSTRACT

Objective: To develop and evaluate a method for quantitative evaluation of *yin* and *yang* (*yin* and *yang* scores) in human subjects for the purposes of research. This method aims to classify subjects into groups allowing future quantitative testing of key research questions such as: do different groups of patients respond differently to acupuncture treatments or Chinese herb formulas?

Methods: In a pilot study of inter-rater reliability, 12 volunteers were each successively interviewed and examined by 6 acupuncturists on the same day. Each acupuncturist gave each volunteer a score for *yin* and a score for *yang* on a scale of -10 to $+10$, zero representing a “balanced” score. Acupuncturists were blinded to each other’s scores.

Results: Overall mean (\pm standard deviation [SD]) *yin* and *yang* scores were -1.86 ± 0.90 and -0.68 ± 1.23 respectively. Intraclass correlations (ICCs) associated with a single acupuncturist’s ratings were 0.35 (*yin*) and 0.36 (*yang*). ICC’s for subject’s mean scores based on the six acupuncturists were 0.77 (*yin*) and 0.78 (*yang*). Significant differences in mean scores across subjects were detected for *yin* ($p < 0.001$) and *yang* ($p < 0.001$) (repeated-measures analysis of variance [ANOVA]) based on the multiple acupuncturists’ ratings.

Conclusion: These results indicate that (1) *yin* and *yang* can be quantified in a reliable manner, but evaluation by multiple acupuncturists is necessary to obtain a reliable score; (2) *yin* and *yang* scores can be used to group individuals for the purposes of statistical analysis. Further evaluation of *yin* and *yang* scores in a greater number and wider variety of patients will be needed to evaluate the potential usefulness of this measurement tool in acupuncture clinical trials and basic physiologic research.

INTRODUCTION

Clinical trials of acupuncture increasingly group research subjects according to Traditional Chinese Medicine (TCM) diagnoses (Hammerschlag, 1998; Schnyer and Allen

2001). This is generally considered desirable because it increases ecologic, clinical, and scientific validity by ensuring that treatment protocols match interventions that are actually consistent with clinical practice. TCM diagnoses, however, are not standardized and tend to be quite complex.

¹Department of Neurology, University of Vermont, Burlington, VT.

²Department of Medicine Biostatistics, University of Vermont, Burlington, VT.

³Acupuncture Vermont, South Burlington, Vermont.

⁴Ancient Roots Traditional Chinese Medicine, White River Junction, Vermont.

⁵Center for Health Studies, Group Health Cooperative, Seattle, WA.

⁶Department of Psychiatry, University of Vermont, Burlington, VT.

⁷Osher Institute, Harvard Medical School, Boston, MA.

Practitioners use a mixture of reasoning based on the “four diagnostic methods” (inspection, listening and smelling, inquiring, and palpation), clinical judgment and intuition to guide this diagnostic process (Kaptchuk, 2000). Studies in which several practitioners examined the same subjects showed considerable variability of diagnosis across practitioners (Hogaboom et al., 2001; Zell et al., 2000). The reliability of TCM diagnoses, therefore, is an important issue in the design of acupuncture clinical trials.

One approach to standardize TCM diagnosis has been to develop algorithms allowing subjects with a specific Western condition to be assigned to a detailed TCM diagnosis (Schnyer 2002a). For example, in patients with major depression, many Chinese medical categories were found, a common diagnosis being “Qi stagnation transforming into heat with dampness accumulation” (Schnyer et al., 2002b). In this study, we propose an alternative approach, which is to group research subjects into broad categories based on essential core components of the TCM evaluation such as *yin* and *yang*, rather than detailed TCM diagnoses. Potential advantages of such an approach are: (1) it may be easier for acupuncturists to agree on fundamental characteristics than on a more detailed evaluation and (2) because of the central importance of *yin* and *yang*, this method may capture enough important elements of TCM diagnosis to make it a valuable research tool.

Yin and *yang* incontestably stand at the core of traditional Chinese natural science, medicine and acupuncture. Quoting a leading acupuncture textbook:

The root cause of the occurrence and development of disease is considered to be an imbalance between *yin* and *yang*. For this reason, however complicated and changeable the clinical manifestations may be, with a good command of the principles of *yin* and *yang*, we may grasp the key linking elements and analyze them effectively. *Yin/yang* is the basis for the differentiation of syndromes by the eight principles, namely *yin*, *yang*, interior, exterior, cold, heat, deficiency and excess. In this way, complicated clinical situations can be simplified, and a correct diagnosis given (Cheng, 1987).

We therefore propose that *yin* and *yang* can be used to evaluate a basic core aspect of an individual’s physiology and pathophysiology.

Yin and *yang* traditionally are illustrated in acupuncture textbooks as bar graphs representing the “relative amount” of *yin* versus *yang*, as well as the “absolute amount” of each relative to a zero point representing a healthy or balanced state (Fig. 1) (Cheng, 1987; Kaptchuk, 2000; Maciocia, 1989; O’Connor and Bensky, 1981). Although the traditional bar graphs pictorially represent the concept of “*yin* greater than *yang*” or “*yang* greater than *yin*,” these concepts to date have not been expressed numerically. The importance of assigning numerical values or “scores” to *yin*

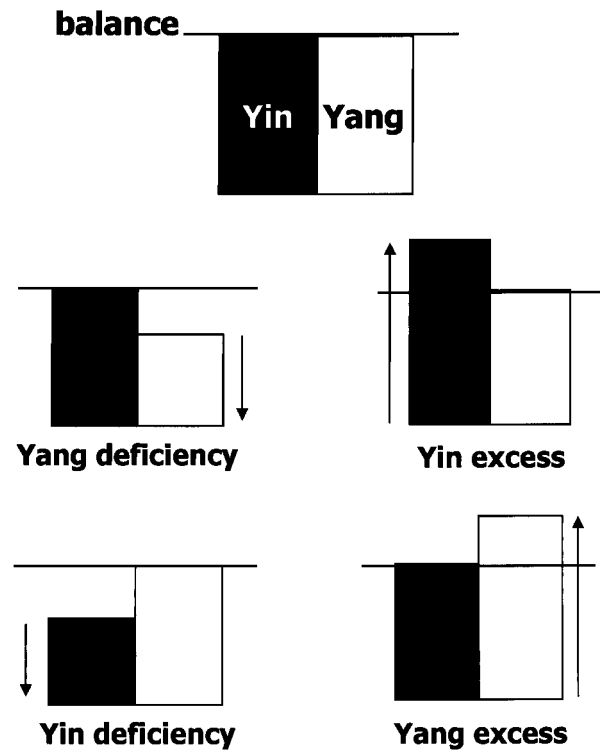


FIG. 1. Typical representation of *yin/yang* patterns in Traditional Chinese Medicine texts. The height of each bar represents the relative amount of *yin* and *yang* relative to a line representing “balance.” *Yin* Deficiency and *Yang* Excess both are characterized by a preponderance of *yang* over *yin*, while *Yang* Deficiency and *Yin* Excess both are characterized by a preponderance of *yin* over *yang*.

and *yang* therefore is that such scores would permit mathematical manipulation of these entities and provide quantitative parameters for statistical analysis.

As a first step to evaluating the potential usefulness of *yin* and *yang* scores in TCM research, this pilot study aimed to test the hypothesis that a numerical score can reliably be assigned to *yin* and *yang*. We asked six acupuncturists to assign each of 12 human subjects a score for *yin* and a score for *yang* on a scale of -10 to $+10$, zero representing a “balanced” score. Importantly, the goal of this study was not to determine what criteria from the history and physical examination are associated with specific *yin* and *yang* scores. We were only interested in knowing how close the agreement would be between acupuncturists and therefore deliberately did not instruct acupuncturists to use a standardized set of questions and examinations. We wanted the acupuncturists to feel free to use whatever method they used in their normal practice.

We proposed that allowing practitioners to use their entire diagnostic skills (including intuition) to perform these evaluations would have the advantage of preserving an important part of the diagnostic evaluation. We hypothesized that acupuncturists would agree on their assessment of *yin* and *yang*, even though they might have used different diagnostic strategies to perform this assessment.

TABLE 1. SUBJECT CHARACTERISTICS

Subject	Age	Past medical history	Current symptoms	Current medication
A	53	Parkinson's disease	Stiffness, slowness of movement	Amantadine, atenolol
B	76	Coronary artery disease, stroke	Pain right leg and hip	HCTZ, Atorvastatin
C	49	None	Headache, neck pain	None
D	56	Bladder surgery, carpal tunnel	Knee pain, urinary incontinence	None
E	40	None	None	Oral contraceptives
F	64	Hepatitis, epidermolysis bullosa	Itchy eyes, insomnia	Aspirin, Ginkgo
G	45	Herpes zoster	Insomnia	None
H	35	Lumbar laminectomy	Back pain	None
I	19	Depression, allergic rhinitis	Runny nose, itchy eyes	Bupropion
J	46	Fractured heel	Back pain	None
K	48	Synovial cysts	Hot flashes, headaches	Black cohosh
L	46	Fractured cervical vertebrae	None	None

HCTZ, hydrochlorothiazide.

METHODS

Twelve (12) adult human volunteers were recruited for participation in the study. There were no exclusion criteria. Subject's ages, past medical history, current symptoms and medications are shown in Table 1. While none of the subjects was "normal," our group of subjects was on the whole relatively healthy, as defined by the absence of acute or chronic illness not controlled by medication. Six licensed acupuncturists currently practicing in Vermont also were recruited. All acupuncturists were Diplomates of the National Certification Commission for Acupuncture and Oriental Medicine (NCCAOM) and had at least 5 years experience practicing acupuncture. All of the acupuncturists practiced "TCM style" acupuncture; however, their backgrounds and training were quite diverse, representing five different schools of acupuncture. Three of the six acupuncturists also held Chinese Herbal Medicine certification from the NCCAOM.

The study took place at the University of Vermont General Clinical Research Center at Fletcher Allen Health Care. Volunteers read and signed an informed consent protocol approved by the University of Vermont Institutional Review Board. Each subject was successively interviewed and examined by the six acupuncturists on the same day. Subjects were divided into two groups of six. The first group was assessed from 8:30 AM to 12:00 PM and the second group from 12:30 PM to 4:00 PM. On both the morning and afternoon, the six subjects each were placed in their own examining room. The six acupuncturists individually rotated from room to room, and were given 30 minutes to assess each subject and write down a *yin* score and a *yang* score before moving on to the next subject.

Each acupuncturist gave each subject a score for *yin* and a score for *yang* on a scale of -10 to +10, zero representing a balanced score. Even though the assignment of numerical *yin* and *yang* scores is not part of a normal TCM evaluation, we felt that acupuncturists would quickly become comfortable with doing so, by first visualizing and

drawing the familiar bar graphs and then simply quantifying the height of the bars. Acupuncturists thus were provided with an empty graph for each patient onto which they drew a bar for *yin* and a bar for *yang* (Fig. 2), then recorded the corresponding *yin* and *yang* scores. Acupuncturists were only required to hand in a *yin* score and a *yang* score for each subject, and were not asked to explain what criteria

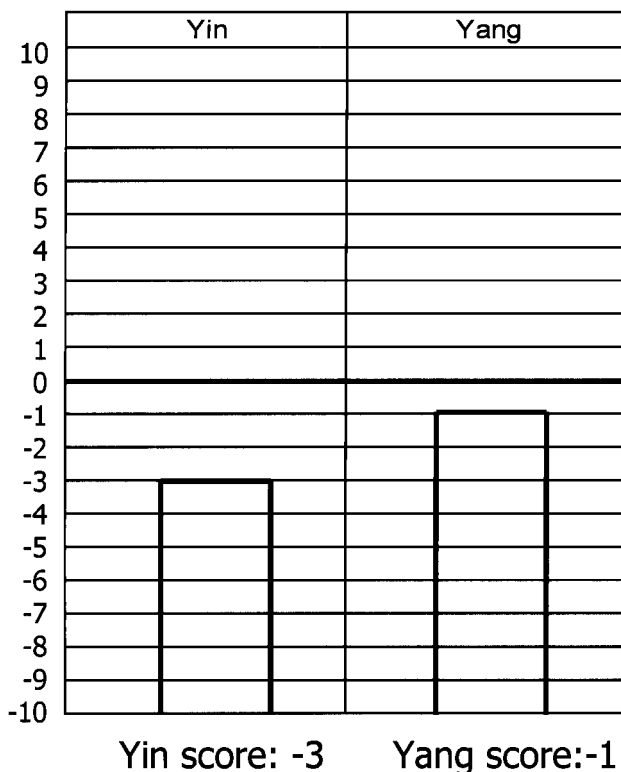


FIG. 2. Example of *yin* and *yang* score sheet as filled out by each acupuncturist. For each subject, acupuncturists were asked to draw a bar for *yin* and a bar for *yang* and then to write down the corresponding numerical scores.

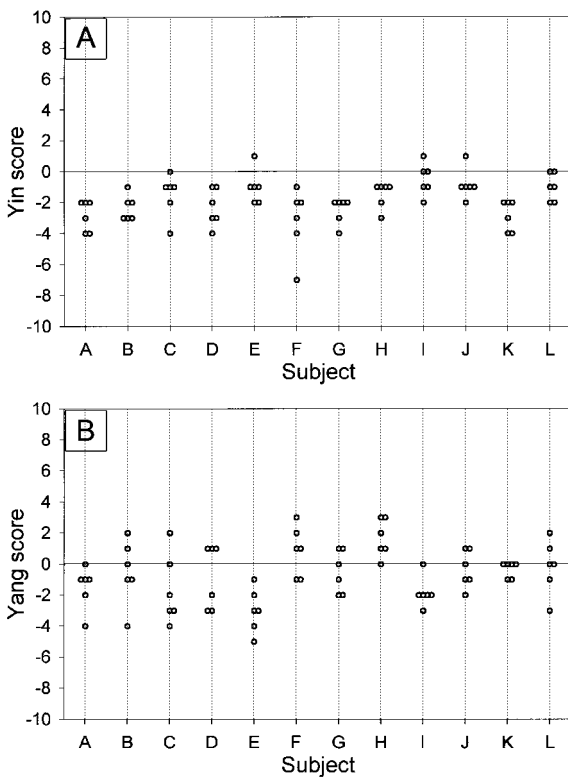


FIG. 3. Raw scores for *yin* (A) and *yang* (B) in 12 human subjects. Each column corresponds to one subject and each circle corresponds to an individual acupuncturist's score for that subject. Scores were rated on a scale of -10 to $+10$, zero representing a "balanced" score.

they had used to arrive at each score. Acupuncturists were blind to each other's scores.

Repeated-measures analyses of variance (ANOVA) were used to test for differences in mean *yin* and *yang* scores across subjects based on the multiple acupuncturists' ratings. The intraclass correlation coefficient (ICC) was computed to measure the reliability associated with *yin* and *yang* scores based on a single acupuncturist and the mean scores across multiple acupuncturists (Winer, 1972). The reliability of a measurement concerns the extent to which this measurement yields the same result on repeated trials, and ICC is the preferred statistic to estimate the reliability of such measurements (Bartko, 1966; Hazzard, 1958). In the case of this study, the measurement is a score given by multiple raters, which is analogous to a score given by multiple judges in a sports competition. If, as expected, the scores vary somewhat from one rater to another, one may not want to rely on the score of a single rater. Multiple raters therefore will be necessary to obtain a reliable score, and the ICC can give an indication of how many raters will be needed. The ICC can be interpreted as the ratio of the true variance (i.e. the variance that truly exists among subjects) to the total variance observed in the scores. This total observed vari-

ance is itself composed of: (1) the true subject-to-subject variability and (2) the additional variability contributed by the raters' inability to replicate each other, either systematically (e.g., one rater consistently scoring higher than another) or randomly. One useful interpretation of the ICC is that its square root represents a hypothetical upper bound for the correlation coefficient between a given measure and other measured attributes. ICCs of approximately 0.80 or greater are considered to represent reasonable reliability. How many raters are necessary to obtain a reliable score can be determined by calculating the ICC for the mean score based on increasing numbers of raters using the Spearman-Brown prophecy formula (Winer, 1972). This allows determination of what would have been the minimum number necessary to achieve acceptable reliability ($ICC = 0.08$). In this study, therefore, the ICCs for *yin* and *yang* scores were computed based on one to six acupuncturists.

RESULTS

Raw scores for *yin* and *yang* are shown in Figure 3. Overall mean (\pm standard deviation [SD]) *yin* and *yang* scores were -1.86 ± 0.90 and -0.68 ± 1.23 , respectively. Estimated ICC associated with increasing numbers of acupuncturists are shown in Table 2. ICCs for a single acupuncturist's ratings were 0.35 (*yin*) and 0.36 (*yang*), while ICCs for subjects' mean score based on all of the six acupuncturists ratings were 0.77 (*yin*), 0.78 (*yang*). Significant differences in mean scores across subjects were detected for *yin* ($p < 0.001$) and *yang* ($p < 0.001$), (repeated-measures ANOVA) based on the multiple acupuncturists' ratings.

DISCUSSION

We report the results of a pilot study testing the inter-rater reliability of a new method for quantitative TCM evaluation. Three important points need to be stressed from the outset. First, this method is intended solely for the purposes of research. Clearly, reducing a TCM evaluation to two numbers eliminates much of the rich complexity inherent in a full TCM diagnosis. We are therefore not suggesting that *yin* and *yang* scores should be used in clinical practice. Sec-

TABLE 2. ESTIMATED RELIABILITY (ICC) OF MEAN *YIN* AND *YANG* SCORES BASED ON VARYING NUMBER OF ACUPUNCTURISTS

Number of acupuncturists	1	2	3	4	5	6
<i>Yin</i>	0.35	0.52	0.62	0.69	0.73	0.77
<i>Yang</i>	0.36	0.54	0.63	0.70	0.74	0.78

ICC, intraclass coefficient correlation.

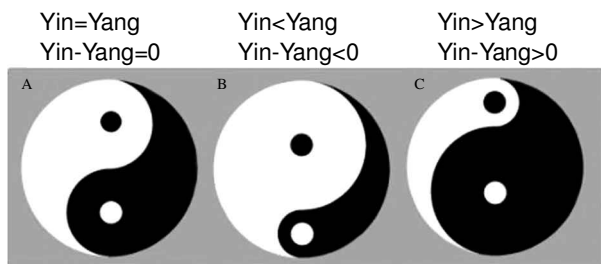


FIG. 4. A pictorial representation of the concept of “yin minus yang” ($yin - yang$). A value of 0 for $yin - yang$ corresponds to balance (A). Negative values for $yin - yang$ correspond to a preponderance of yang over yin (B), while positive values correspond to a preponderance of yin over yang (C).

ond, the method presented here aimed to capture an essential core component of TCM diagnosis. This core component, however limited, is based on TCM and therefore should be more appropriate for TCM research than would categorization based on a Western diagnostic framework. Third, the small number of subjects presented in this paper implies that further studies will be needed to fully evaluate the method’s applications in a variety of research settings. We believe that the novelty of the method and the data presented will challenge the field of TCM, and at the least will encourage discussion and further exploration of quantitative TCM evaluation.

A key question in TCM research is whether patients placed in different groups based on TCM criteria respond differently to different types of acupuncture treatments (e.g., different needle manipulation methods) or Chinese herb formulas. The results of this study indicate that *yin* and *yang* scores can be determined in a reliable manner, and therefore may become useful quantitative research tools. Moreover, mathematical manipulation of *yin* and *yang* scores allows calculation of quantities such as “yin minus yang” ($yin - yang$) and “yin plus yang” ($yin + yang$) that are directly related to important TCM concepts. $yin - yang$ (or the relative difference between *yin* and *yang*) can be thought of as representing which of *yin* or *yang* predominates over the other (Fig. 4). $yin + yang$, on the other hand, can be thought of as representing the “total combined amount” of *yin* and *yang*, representing TCM concepts of Excess and Deficiency. In this study, plotting each subject’s mean ($yin + yang$) and ($yin - yang$) values respectively as x and y Cartesian coordinates yielded four groups A, B, C, and D (Fig. 5). Eight subjects fell into group C ($yin - yang < 0$; $yin + yang < 0$), three in group A ($yin - yang > 0$, $yin + yang < 0$), one in group D ($yin - yang < 0$; $yin + yang > 0$) and none in group B ($yin - yang > 0$; $yin + yang > 0$).

Figure 6 illustrates how *yin* and *yang* scores may overlap with TCM diagnoses. For example, patients with a simple diagnosis of *yin* Deficiency would cluster into group C because *yang* would predominate over *yin* in the context of

Deficiency. A patient with a more complex diagnosis (such as Deficient Kidney *Yin* and Excess Liver *Yang*) would likely fall into either group C or group D (because *yang* would be relatively greater than *yin*). Whether this individual would fall into group C or D would depend on whether the *Yin* Deficiency was more pronounced than the *Yang* Excess, or vice versa. Other types of diagnoses (e.g. Stagnant Liver *Qi* or Dampness Accumulation) may also cluster in different groups: for example, Stagnant Liver *Qi* is often associated with a preponderance of *yang*, and Dampness Accumulation with a preponderance of *yin*. Future studies also may address, for example, whether patients given a diagnosis of either Stagnant Liver *Qi* or Liver *Yang* Rising Upward would cluster in different parts of a given quadrant in Figure 5. The salient point here is that *yin* and *yang* scores should not be considered substitutes for a detailed TCM diagnosis. Rather, the scores provide a way to classify patients in a simple manner that is easily amenable to quantitative analysis. Because of its simplicity and limited scope, the method presented in this paper may indeed be more replicable than TCM diagnosis, although testing this hypothesis was not the focus of this study.

Yin and *yang* scores may allow testing of specific treatment hypotheses in clinical trials. For example, some basic acupuncture or Chinese herbal treatments could be administered to patients placed in different groups based on *yin* and *yang* score categories (such as categories A, B, C, and D in Fig. 5). While these diagnostic categories would not

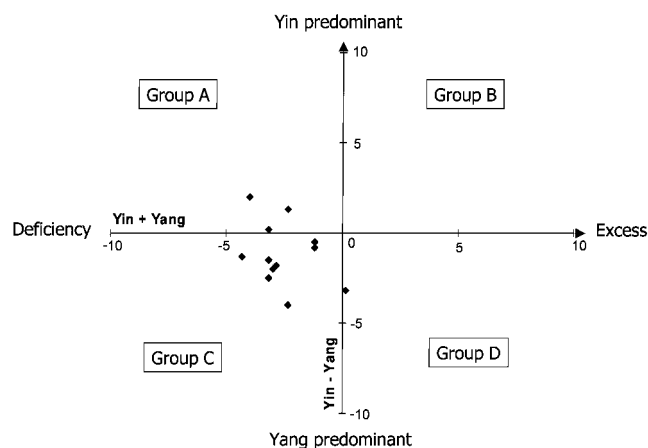


FIG. 5. Cartesian plot of mean values for $yin + yang$ and $yin - yang$ for 12 human subjects. $yin + yang$ is represented on the x axis, and $yin - yang$ on the y axis. Values on the x axis represent the “total amount” of *yin* and *yang* taken together: positive values (groups B and D) thus represent total amounts of *yin* and *yang* exceeding 0 (Excess), while negative values (groups A and C) represent total amounts less than 0 (Deficiency). Values on the y axis, however, represent whether *yin* or *yang* predominates over the other. Positive values (groups A and B) correspond to a preponderance of *Yin* over *Yang*, while negative values (groups C and D) correspond to preponderance of *Yang* over *Yin*.

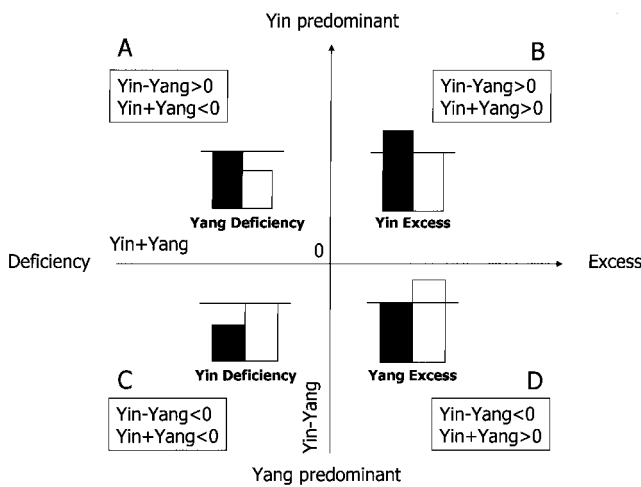


FIG. 6. Overlap of groups based on *yin* – *yang* and *yin* + *yang* with simple Traditional Chinese Medicine diagnoses of *Yang* Deficiency (group A), *Yin* Excess (group B), *Yin* Deficiency (group C), and *Yang* Excess (group D).

be as detailed and individualized as they would be if based on a full TCM diagnosis, the patient grouping nevertheless would be based on basic TCM principles. This approach therefore would be closer to TCM practice than grouping based on Western diagnoses. A plausible hypothesis would be that patients in group A would respond better to warming herbs and tonification of points such as Spleen 3 and GV4, compared to patients in group D.

In addition to its potential usefulness in clinical trials, *yin* and *yang* scores may be relevant to basic physiologic research, allowing testing of hypotheses relating to correspondences between Eastern and Western systems of medicine. For example, does “*Yin* Deficiency” correlate with any specific western physiologic measurements? *Yin* and *yang* have been described as “undergarments,” normally invisible but only showing inappropriately in pathologic situations (Larre et al., 1986). An imbalance between *yin* and *yang* may represent a physiologic pattern, recognizable through signs and symptoms on history and physical examination (Taitano et al., 2002). Although we did not ask the acupuncturists in this study to describe how they arrived at *yin* and *yang* scores, this could be explored in a further study.

Our results suggest that *yin* and *yang* can be reliably quantified, but that multiple acupuncturists are necessary to obtain a reliable score. The lack of reliability of a single acupuncturist should not be viewed as an indication that the acupuncturists performed poorly. Rather, this should be understood as being a consequence of the natural variability that accompanies any kind of subjective assessment. Although in this first study we found that four to six acupunc-

turists were required for ICCs to approach an acceptable range, it is not unreasonable to assume that this number could be reduced in future studies. How many acupuncturists ultimately are required obviously is an important practical question to determine the feasibility of using *yin* and *yang* scores in clinical trials. An important factor influencing the magnitude of the ICC (i.e., the estimated reliability of the measurements) is the amount of true variability that exists in a given group of subjects with respect to *yin* and *yang* (also referred to as the “range of talent” effect). The low variability in *yin* and *yang* scores in our study may have been due to our subjects being relatively healthy with few symptoms (for example, subject A had mild Parkinson’s disease and subject B’s stroke and coronary artery disease were currently asymptomatic). Measurement of *yin* and *yang* scores in less healthy patients with more pronounced symptoms and pathology would be expected to yield improved ICCs. Training of acupuncturists also would be expected to improve ICCs. An interesting observation in this study was that ICCs improved in the six patients tested in the afternoon, compared with those tested in the morning, which suggests a training effect due to practice or familiarity with using the instrument (because no conversation about the instrument took place between the morning and afternoon sessions). Indeed all acupuncturists agreed that they felt increasingly comfortable with the rating process as the day progressed, and most acupuncturists took substantially less than the allotted 30 minutes to arrive at a score in the afternoon session.

A limitation of this study was that although all raters had at least 5 years experience with TCM, they were not identical in terms of education within TCM, since only three of the six were licensed in Chinese herbal medicine. It is therefore possible that ICCs would have been higher if the acupuncturists’ education had been more homogeneous.

In summary, a simple question was posed in this study and the resulting data suggest that, indeed, *yin* and *yang* can be quantified in a reliable manner, and may be useful to classify research subjects into categories for the purposes of statistical analysis. Further evaluation of *yin* and *yang* scores in a greater number and wider variety of patients will be needed to evaluate the potential usefulness of this measurement tool in acupuncture clinical trials and basic physiologic research.

ACKNOWLEDGMENTS

We thank Margery Keasler, L.Ac., Glynn Pellegrino, L.Ac. and Kirk White, L.Ac. for performing the subjects’ evaluations. This study was conducted at the University of Vermont General Clinical Research Center at Fletcher Allen Health Care supported by National Institutes of Health (NIH) Grant RR00109.

REFERENCES

- Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep* 1966;19:3–11.
- Cheng X. Chinese Acupuncture and Moxibustion. Beijing: Foreign Language Press, 1987.
- Hammerschlag R. Methodological and ethical issues in clinical trials of acupuncture. *J Altern complement Med* 1998;4:159–171.
- Hazzard E. Intraclass Correlation Coefficient and Analysis of Variance. New York: Dryden Press Inc., 1958.
- Hogaboom CJ, Sherman KJ, Cherkin DC. Variation in diagnosis and treatment of low back pain by Traditional Chinese medicine acupuncturists. *Complement Ther Health Med* 2001;9:154–166.
- Kaptchuk TJ. The Web That has No Weaver. Understanding Chinese Medicine. Chicago: Contemporary Publishing Group, Inc., 2000.
- Larre C, Schatz J, Rochat de la Vallee E. Survey of traditional Chinese medicine. Paris: L'Institut Ricci, 1986.
- Maciocia G. The Foundation of Chinese Medicine. Edinburgh: Churchill Livingstone, 1989.
- O'Connor J, Bensky D. Acupuncture: A Comprehensive Text (Shanghai College of Traditional Medicine). Seattle: Eastland Press, 1981.
- Schnyer RN, Allen JJB. Acupuncture in the Treatment of Major Depression. A Manual for Practice and Research. London: Harcourt-Brace Churchill Livingstone, 2001.
- Schnyer RN, Allen JJB. Bridging the gap in complementary and alternative medicine: manualization as a means of promoting standardization and flexibility of treatment in clinical trials of acupuncture. *J Alt Compl Med* 2002a;8:623–634.
- Schnyer RN, Taitano K, Allen JJB. Prevalence of Chinese medicine defined patterns in people with major depression. Proceedings of the 9th Annual Symposium of the Society for Acupuncture Research, 2002b.
- Taitano K, Schnyer R, Allen JJB, Manber R, Hitt S. The psychophysiology of *yin* and *yang*. International Scientific Conference on Complementary, Alternative and Integrative Medical Research. April 2002.
- Winer BJ. Single-factor experiments having repeated measures on the same elements. *Statistical Principles and Experimental Design*, New York: McGraw Hill, 1972:261–301.
- Zell B, Hirata J, Marcus A, Ettinger B, Pressman A, Ettinger KM. Diagnosis of symptomatic postmenopausal women by traditional Chinese medicine practitioners. *Menopause*. 2000;7:129–134.

Address reprint requests to:
Helene M. Langevin M.D., L.Ac.
Department of Neurology
University of Vermont
Given C 423
Burlington VT 05405

E-mail: Helene.Langevin.uvm.edu