THE JOURNAL OF ALTERNATIVE AND COMPLEMENTARY MEDICINE
Volume 15, Number 4, 2009, pp. 353–366

DOI: 10.1089/acm.2008.0455

A Review of the Reliability of Traditional East Asian Medicine Diagnoses

Kylie A. O'Brien, Ph.D.,[1] and Stephen Birch, Ph.D.[2]

Abstract

Background: In the practice of forms of traditional East Asian medicine (TEAM) such as Traditional Chinese Medicine and Japanese meridian therapy, treatment follows identification of underlying "patterns of disharmony." However, little is known in an objective sense of the consistency or reliability of diagnosis within TEAM. This is important: If diagnosis is not reliable, there can be less confidence that optimal treatment is received. TEAM systems have their own diagnostic endpoints that are used as evidence of change. If these are to be incorporated into clinical studies, a prerequisite is that they are reliable. Few studies have assessed the reliability of diagnostic data collected during a TEAM examination. The majority have investigated reliability of pulse diagnosis, with results ranging from low to a very good level of agreement. Studies of reliability of tongue diagnosis and other diagnostic data collected in a Chinese Medicine examination suggest considerable variability. In general, studies of reliability of pattern diagnosis and treatment in a range of disorders have not found a high level of reliability. A range of factors may affect reliability.

Objectives: This paper reviews the current knowledge of reliability of TEAM diagnoses, including limitations of studies, and discusses the implications for practice and research and how to improve the current situation.

Introduction

Traditional East Asian Medicine (TEAM) refers to a diverse set of therapies originating in East Asia, including *zhong yi* or Traditional Chinese Medicine (TCM). Within each traditional medicine system, variations of traditional concepts, theories, and models of how the body functions guide how information is collected and organized in order to decide the treatment.[1] Traditional methods of diagnosis and diagnostic classifications are performed principally to inform the root treatment selection (that is, one that targets the underlying cause of the illness).[2,3] Thus, diagnosis is principally a pointer to the treatment, rather than an "objective" description of what is wrong with the patient,[1] unlike in Western medicine.

Relatively little is known about how consistent diagnosis is within various systems of TEAM. This is important since if diagnosis is not consistent, treatment is unlikely to be consistent. This has important implications for clinical practice and education. In addition, scientific research is being conducted into TEAM systems and therapies. TEAM systems have their own diagnostic variables that a practitioner uses as evidence of change within the patient; however, this is only justifiable if such variables are consistent.

Reliability, also called reproducibility, refers to the consistency of information, that is, the extent to which similar information is acquired when a measurement is conducted more than once.[4] The degree of agreement between observers conducting the same measurement is termed "inter-rater reliability." The degree of agreement when the same observer repeats the observations is called "test–retest reliability" or "intra-rater reliability." One statistical method, calculation of the "κ coefficient," is commonly used to measure the level of agreement beyond that expected by chance and provides a measure of inter-rater reliability. An interpretation of κ values in terms of level of agreement, proposed by Landis and Koch,[5] is set out in Table 1.

Calls for use of reliability studies of TEAM diagnostics have been made since the late 1980s.[6–11] with few studies conducted. This paper reviews the current knowledge of reliability of TEAM diagnoses and discusses the implications for practice and research and how to improve the current situation.

[1]Faculty of Health, Engineering and Science, Victoria University, Melbourne, Australia.
[2]Stichting (Foundation) for the Study of Traditional East Asian Medicine (STEAM), Amsterdam, The Netherlands.

Table 1. Interpretation of κ Values[5]

κ value	Level of agreement
<0.0	Poor
0.0–0.20	Slight
0.21–0.40	Fair
0.61–0.80	Substantial
0.81–1.00	Almost perfect

Review of the Current Literature

The first studies of reliability of diagnosis in TEAM were conducted in the 1960s in Japan,[12–15] with increasing interest in Western countries occurring in the late 1980s.[16] Reliability studies have been conducted on pulse diagnosis (see footnotes[*†‡]),[17–21] tongue diagnosis (see footnote[§]),[21,22] other diagnostic judgments, [21] and overall TEAM pattern diagnosis (see footnotes[*†]),[23–32] with variable results. We will briefly discuss and summarize the findings of the more important studies. The purpose of this review was not to comprehensively critique individual studies relating to reliability of TEAM. Rather, it was to present a summary of findings of reported studies and point out various issues relating to reliability of TEAM, in some cases using particular studies as examples.

Methodology

A literature search was conducted of studies published in English on the reliability of TEAM. Key search terms included "reliability," "repeatability," "inter-rater reliability," and "Chinese medicine," "Traditional Chinese Medicine," "TCM," "tongue diagnosis." "Chinese medicine syndrome diagnosis," and "pulse diagnosis." The major electronic databases used were PubMed, Medline and "Google." As papers were collected, the bibliographies were scrutinized for additional papers relating to reliability, which were then sourced. Studies relating to Japanese forms of TEAM had previously been sourced by Stephen Birch as part of his doctoral thesis.*

Reliability of tongue diagnosis

Few studies have investigated the reliability of tongue diagnosis. Two studies using practitioners[22] and students[§] as observers viewing slides of tongues found a low level of inter-rater reliability for most characteristics of tongue diagnosis except where there were dichotomous response choices.[22] One study also found a low level of intra-rater reliability.[22] Two other studies utilized actual patients. Dang and Zaslawski[17] reported a low level of agreement between 6 TCM practitioners examining a patient with cystic fibrosis. O'Brien and colleagues[21] investigated the reliability of clinical data collected by 3 TCM practitioners in 45 patients with hypercholesterolemia and found the level of agreement for tongue coating and tongue body characteristics ranged from "fair" to "moderate" and "slight" to "moderate," respectively. Several potential sources of variability were acknowledged.[21] Details of the studies are shown in Table 2.

Reliability of pulse diagnosis

Several studies have investigated the inter-rater reliability of pulse diagnosis. Cole conducted a series of small studies that examined reliability and validity of pulse diagnosis (see footnote[‖]) and found poor reliability.[11]

Kass[16] examined agreement on pulse diagnosis between two TCM practitioners, 1 of whom performed manual pulse diagnosis and the other of whom used a machine that he had developed that had previously been calibrated to the findings of the first practitioner.[33] Though a high degree of agreement was claimed between the practitioner and the device (agreement in 87/110 cases, $p < 0.0001$),[16] no data were presented that established the reliability of the first practitioner's judgments in comparison to other practitioners, raising questions about the usefulness of this study.

The results of several studies of inter-rater reliability of pulse diagnosis (see footnotes[*†‡])[17,18,20,21] are summarized in Table 3. Not all studies used formal statistical tests for reliability, and a variety of approaches have been used to analyzing and reporting data. Birch's studies (see footnotes[*†]) were conducted in the Toyohari Meridian Therapy system, whereas others investigated pulse diagnosis within TCM models. The majority of studies investigated reliability of basic pulse qualities such as depth, speed, and strength rather than agreement on the more complex 28 pulses in the TCM literature.

Comparability across studies is difficult due to a range of reasons. Some studies utilized TCM students as observers,[20,34] although the majority used practitioners. Lack of experience is likely to reduce the possibility of agreement, although studies using student observers are still useful to inform educators.

Conclusions of various studies range from a low level of reliability of pulse diagnosis[17] through moderate agreement (see footnote [‡]) to very good agreement.[18,21] Research to date suggests that as level of complexity of pulse detection increases, reliability of pulse diagnosis decreases.[16] Difficulties in pulse assessment have included difficulties of translation and lack of standard pulse terminology, uncertainty in definitions of the 28 recognized pulses,[18,35] and lack of standard pulse taking procedure.[18]

*Birch S. An exploration with proposed solutions of the problems and issues in conducting clinical research in acupuncture [Ph.D. thesis]. Exeter University, 1997.

†Birch S. Preliminary investigations of inter-rater reliability of traditional based acupuncture diagnostic assessments [unpublished manuscript]. 1999.

‡Craddock DS. Is traditional Chinese medical pulse reading a consistent practice? A comparative pilot study of four practitioners [undergraduate independent research project]. Sydney: University of Technology, 1997.

§Rupp N. Tongue diagnosis: An analysis of its reliability as a diagnostic tool [undergraduate independent research project]. Sydney: University of Technology, 1998.

‖Cole P. Acupuncture and pulse diagnosis in Great Britain [unpublished Ph.D. thesis]. University of Sussex, 1975.

Table 2. Inter-rater Reliability Studies of Tongue Diagnosis

Authors	*Subjects studied*	*Observers*	*Study design features*	*Statistical test*	*Results*
Rupp 1998[a]	5 photographs of tongues (2 of which were identical)	26 final year acupuncture students	Students asked to rate the tongue photographs according to tongue body color and shape and presence and color of tongue coating (13 questions in total).	None reported.	A low level of agreement was found on most aspects of tongue diagnosis except for presence of cracks.
Dang & Zaslawski 1998[17]	One subject with cystic fibrosis	6 TCM practitioners (4 experienced practitioners and 2 recent graduates)	Practitioners asked to diagnose, describe pulse and tongue, and to suggest a treatment principle and two acupuncture point prescriptions.	Not reported.	Low level of reliability for pulse and tongue diagnosis.
O'Brien et al. 2008[21]	45 healthy Australians with hypercholesterolemia	3 TCM practitioners		κ coefficient.	Level of agreement between 3 practitioners for tongue body characteristics: constitution 48%, κ 0.31; size 25% (κ 0.20); color 13% (κ 0.07); presence/absence of papillae 37% (κ 0.16). Level of agreement between at least 2 practitioners: constitution 48%, κ 0.31; size 90% (κ 0.73); color 56% (κ –0.19); presence/absence of papillae 37% (κ 0.16). Level of agreement between 3 practitioners on tongue coating characteristics: quality 34% (κ 0.31); thickness 27% (κ 0.22); color 43% (κ 0.41). Level of agreement between *at least 2* practitioners on tongue coating characteristics: quality 95% (κ 0.90); thickness 95% (κ 0.87); color 91% (κ 0.90).
Kim et al. 2008[22]	10 Power Point slides of tongues	30 TCM practitioners	Participants completed two questionnaires: a standard color slide questionnaire to test color discrimination, and a tongue slide questionnaire with 11 questions about tongue body and coating characteristics. The study tested inter-rater and intra-rater reliability. Criterion agreement levels were set at ≥80%.	None reported.	Level of agreement for most tongue characteristics was low. Reliability criterion overall only achieved on 5% of occasions. High level of agreement (≥80%) for presence of tongue coating (yes/no choice) in all 10 slides and for presence of crack (yes/no choice) in 5/10 slides. Highest intrapractitioner agreement achieved for two questions offering dichotomous choice (yes/no): presence of cracks and presence of coat (≥80% agreement for 29/30 and 20/30 practitioners, respectively). Levels of test–retest agreement were poor for most other tongue characteristics, where the number of response choices was between 5 and 8.

[a]Rupp N. Tongue diagnosis: An analysis of its reliability as a diagnostic tool [undergraduate independent research project]. Sydney: University of Technology, 1998.

Table 3. Iter-rater Reliability Studies of Pulse Diagnosis

Authors	*Subjects*	*Observers*	*Study design features*	*Statistical test*	*Results*
Birch 1997[a]	9 subjects	5 Meridian therapy practitioners	Case history was conducted by participating practitioners together then each observer rotated separately through a room. Subject sat quietly and still to minimize potential variability due to activity. Communication between practitioners prohibited. Each basic pulse quality (pulse depth, speed, strength) rated on an ordinal 1–5 scale.	Spearman rank correlation test/coefficient (Cohen's κ would not compute because all 5 points on each ordinal scale were not used by the practitioners).	Inter-rater agreement is given in a range: Pulse depth - $r = -0.04$ (no correlation) to 0.75 (substantial correlation) (average 0.43). Pulse rate - $r = 0.04$ (poor correlation) to $r = 0.69$ (substantial correlation) (average 0.38). Pulse strength - $r = -0.004$ (no correlation) to $r = 0.93$ (almost perfect correlation) (average 0.45).
Birch 1999[b]	43 subjects	2 Meridian therapy practitioners	Extension of above study.	κ coefficient and when it would not compute, the Spearman rank correlation test/ coefficient (see above).	Inter-rater agreement significant for pulse rate ($K = 0.29$, $T = 3.17$, $p < 0.01$). Nonsignificant for pulse depth ($K = 0.02$, $p = 0.82$). For pulse strength the Spearman rank correlation was significant ($r = 0.84$, $p < 0.001$).
Craddock 1997[c]	8 subjects	4 TCM practitioners		Not reported.	Moderate agreement (63%) between 4 practitioners in pulse diagnosis. Average intrapractitioner level of consistency 57%. As subtlety and complexity of the pulse category increased, level of agreement between practitioners decreased.
Dang & Zaslawski 2003[17]	One subject with cystic fibrosis	6 TCM practitioners (4 experienced practitioners and 2 recent graduates)	Practitioners asked to diagnose, describe pulse and tongue, and to suggest a treatment principle and two acupuncture point prescriptions.	Not reported.	Low level of reliability for pulse and tongue diagnosis.
King et al. 2002[18]	66 subjects in Collection 1 (initial data collection); 30 subjects in Collection 2 (replication collection). Subjects recruited from University staff and students and general population.	2 TCM practitioners	Operational definitions and standardized manual palpation method developed based on literature review of pulse definitions and pulse-taking methods and practical test–retest method. Definitions given for pulse depth, width, force, relative force, rhythm and location, pulse occlusion.	16 pulse categories relating to length, depth, force, relative force, ease of occlusion, and rhythm. Did not assess pulse speed. Inter-rater agreement measured as percentage agreement between 2 assessors. χ^2 test.	Mean inter-rater agreement of 80% in Collection 1 and Collection 2. Greater than 70% agreement achieved for 13 of the 16 categories and greater than 80% agreement achieved for 10 categories in Collection 1. Levels of at least 80% obtained for 11 of 16 data categories in Collection 2.

			Standard assessment form to record data. Data recorded as categorical variables, assessors required to choose from limited range of answers.		
O'Brien et al. 2008[21]	45 patients with hypercholesterolemia who were otherwise healthy	3 TCM practitioners	Part of a study on inter-rater reliability on data collection using three of the four diagnostic methods in a Chinese medicine examination. Standard assessment form to record data. Data recorded as categorical variables, assessors required to choose from limited range of answers.	κ coefficient.	Agreement between 3 practitioners was "slight" for pulse location (24%, κ 0.15) and "fair" for pulse force (37%, κ 0.29). Agreement between at least two practitioners: "almost perfect" for pulse location (100% agreement, κ 1.00) and pulse force (97% agreement, κ 0.86). Agreement on pulse speed assessed between 2 practitioners: "moderate" agreement (75%, κ 0.63). Comparison of pulse speed measured by Practitioner 1 using an objective method (Dynamap Blood Pressure Instrument) and traditional method (counting beats/breath): 89% agreement between Practitioner 1 and Practitioner 2 (κ 0.84) and 81% agreement between Practitioner 1 and Practitioner 3 (κ 0.72), respectively.
Walsh et al. 2001[20]	6 subjects divided into 2 groups (of 3 subjects each)	TCM students randomly assigned to 1 of 2 groups: 35 at baseline, 29 at end of 14 weeks, 20 a year later.	Pulse discrimination examined at 3 time-points: Collection 1 (baseline), Collection 2 (end of 14 weeks, completion of pulse-taking classes) and Collection 3 (1 year later). Measurements taken within 2–3 hours. Subjects positioned behind screen. Standardized form to record data.Assessed inter-rater agreement on pulse speed, depth, strength, and length (paper details definitions of each of these).	12 pulse characteristics examined on 6 subjects. Level of agreement for each collection phase above that expected by chance calculated using χ^2 test (significance set at $p = 0.05$).	Frequency of significant levels of pulse discrimination in Collection 1, 2, and 3 respectively: 25%, 31%, and 17% occasions, respectively. Significant difference in discrimination between the *cun* and *chi* positions for pulse characteristic full/empty: *chi* position highest (61% left, 55% right wrist) then *guan* then *chi*. Across the three collections, the lowest levels of pulse discrimination were recorded in Collection 3. Agreement more frequent between observers when palpating pulse of female subjects though did not reach statistical significance.

[a]Birch S. An exploration with proposed solutions of the problems and issues in conducting clinical research in acupuncture [Ph.D. thesis]. Exeter University, 1997.
[b]Birch S. Preliminary investigations of inter-rater reliability of traditional based acupuncture diagnostic assessments [unpublished manuscript]. 1999.
[c]Craddock DS. Is traditional Chinese medical pulse reading a consistent practice? A comparative pilot study of four practitioners [undergraduate independent research project]. Sydney: University of Technology, 1997.
TCM, Traditional Chinese Medicine.

Reliability of other clinical observations

Few studies have assessed reliability of clinical diagnostic data other than tongue appearance and pulse. Matsumoto[14] examined the test–retest reliability of abdominal diagnosis. In three separate studies, six, four, and six blindfolded observers completed abdominal diagnosis in 8, 12, and 10 subjects, respectively. Each subject was examined twice by each observer in random order. The percentage of repeated findings (test–retest) ranged from 25% to 90% across the 16 observers (average 61%). Level of agreement between the two highest scoring observers was then conducted and found to range from 25% to 70%. Matsumoto[14] concluded that abdominal palpation appeared to be a reliable diagnostic procedure, although no formal statistical analyses were conducted.

O'Brien and colleagues' study in hypercholesterolemic Australians found that level of agreement between 3 practitioners varied from slight (e.g., color around eyes), fair (e.g., color of complexion), moderate (voice strength), substantial (e.g., character of breath sounds), to almost perfect (presence of spirit).[21]

Reliability of pattern diagnosis in TEAM

Researchers have investigated the reliability of pattern diagnosis in Japanese meridian therapy (see footnotes *†)[12,13,15] and more recently, in TCM.[23–32] The majority of the TCM studies used study designs that require the practitioners to choose from a list of possible syndromes (although some made provision for additional diagnoses), with one study leaving this completely open-ended, supplying no predefined list of syndromes.[27]

Test–retest reliability of pattern diagnosis

Three studies have assessed test–retest reliability of pattern diagnosis based on pulse diagnosis alone, all within the Japanese meridian therapy system (see footnote *).[12,13] These are shown in Table 4. In Debata's study, the percentage of repeated diagnoses averaged 52% over three substudies.[12] Only the results for 6 of the 17 practitioners were considered significant on test–retest reliability and only 2 of the practitioners showed significant inter-rater agreement.[12]

Kurosu similarly conducted a study of test–retest reliability of pattern diagnosis on the basis of only pulse diagnosis in 40 subjects divided into two groups (those who were healthy and those with a medical condition).[13] The mean percentage of repeated findings was 53% in healthy patients and 60% in diseased patients, with an average test–retest percentage of 56% pooling the results. Kurosu assumed that if pattern selection by pulse diagnosis is valid, there should be a higher test–retest percentage in diseased patients compared to healthy patients. Since his results did not bear this out, he concluded that pattern selection by pulse diagnosis is not valid. However, this conclusion is questionable. First, no formal statistical analyses were conducted, and second, one does not normally perform pattern diagnosis on the basis of radial pulse diagnosis alone.[15] The studies therefore did not assess actual clinical practice.

In contrast, Birch examined test–retest both on the basis of pulse diagnosis alone and on the basis of two diagnostic features (pulse and abdominal diagnoses) using 1 practitioner (see footnote *). He found that the reliability of the diagnostic pattern through pulse assessment alone was poor but when two diagnostic factors were examined, the Kendall coefficient of concordance, though relatively low (0.1053), approached statistical significance.

These results need to be interpreted with some caution since in studies where only 1 observer is used, it is not possible to separate variability due to the subjects and variability due to the observer. It appears that the combination of diagnostic assessment methods increases the reliability of the overall diagnostic conclusion, in keeping with Ogawa's claim (see next section),[3] and not dissimilar to what occurs in general medical practice when the results of two different tests are seen together.[36]

Inter-rater reliability of pattern diagnosis

Several studies have investigated the inter-rater reliability of pattern diagnosis within Japanese meridian therapy (see footnotes *†)[15] and within TCM[23–32] with variable results (Table 5).

Two (2) studies[28,32] demonstrated increased agreement between practitioners following training sessions focusing on diagnostic procedures and reaching consensus respectively that were incorporated into studies.

Results from Birch's studies in Japanese meridian therapy (see footnote *) suggest, as claimed by Ogawa,[15] that better results are to be expected when diagnostic methods are used together, for example, pulse diagnosis combined with abdominal diagnosis (as in clinical practice) rather than in isolation.

Reliability of the eight guiding principles

Few studies have assessed the reliability of the eight guiding principles. Hogeboom and colleagues' study of 6 practitioners who examined 6 patients with low-back pain found that all practitioners reported the eight guiding principles as useful in at least 50% of patients, with discrimination between the principles of excess or deficiency considered the most important.[24] There was little agreement on which principles were important for which patients, but no statistical analysis was conducted.[24]

O'Brien and colleagues assessed the inter-rater reliability of the eight guiding principles in 3 practitioners who examined 45 Australians with hypercholesterolemia.[27] When agreement between all 3 practitioners was considered, the level of agreement ranged from "almost perfect" agreement for location (interior/exterior) to "fair" (*yin/yang* summary principle) to only "slight" (excess/deficiency, heat/cold principles).

Reliability of TCM treatments

A small number of studies have investigated the reliability of TCM treatments.[17,23–25,28,30] These are summarized in Table 5. The interpretation of the reliability of treatments with acupuncture or herbal medicine requires some caution. Part of the art of and a characteristic of TEAM is an individualized approach to treatment. Therefore, it may not be surprising if, for example, acupuncture points chosen to treat a condition were to vary between practitioners. Sung and

TABLE 4. TEST–RETEST RELIABILITY STUDIES OF PATTERN DIAGNOSIS

Authors	*Subjects*	*Observers*	*Study design features*	*Statistical test*	*Results*
Debata 1968[12]	30 Subjects:substudy 1: 8 subjects substudy 2: 12 subjects substudy 3: 10 subjects	17 Meridian therapy practitioners in groups of 6 (substudy 1), 6 (substudy 2), and 5 (substudy 3)	3 Substudies examining the test–retest reliability of pattern diagnosis based on pulse diagnosis alone. Practitioners were blindfolded, examined each subject twice in random order, and were required to choose from four possible patterns.	No formal statistical analysis.	Percentage of repeated diagnoses: Substudy 1: 25–63% (mean 45.8%) Substudy 2: 8–75% (mean 40.3%) Substudy 3: 50–90% (mean 64%) Overall mean test–retest result 52.3%.
Kurosu 1969[13]	40 subjects, half healthy and half with a medical condition	17 Blindfolded meridian therapy practitioners	4 Substudies examining test–retest reliability of pattern diagnosis based on pulse diagnosis alone. Practitioners were blindfolded and examined each subject twice in random order.	No formal statistical analysis.	Percentage of repeated diagnoses: Healthy patients 43–65% (mean 53%) Diseased patients: 50–75% (mean 60%). Pooling all subjects, average test–retest percentage 55.8%.
Birch 1997[a]	35 subjects	1 Blindfolded meridian therapy practitioner	Test–retest reliability of pattern diagnosis based on pulse assessment alone.	Kendall coefficient of concordance.	$W = 0.0018$ (very low).
Birch 1997[a]	19 subjects	1 Blindfolded meridian therapy practitioner	Test–retest reliability of pattern diagnosis based on a match between radial pulse diagnosis and abdominal palpation. Each subject examined 4 times in random order.	Kendall coefficient of concordance.	$W = 0.1053$: low but approached statistical significance ($p = 0.11$).

[a]Birch S. An exploration with proposed solutions of the problems and issues in conducting clinical research in acupuncture [Ph.D. thesis]. Exeter University, 1997.

colleagues' study in patients with irritable bowel syndrome demonstrated an improvement in level of agreement on treatment with incorporation of a training phase in the study.[28]

Discussion

It can be seen from the studies conducted thus far that there is a variable degree of consistency in TEAM diagnosis from collection of data through pattern diagnosis. This is not unlike Western medicine, where considerable variation in diagnostic methods conducted as part of a Western physical examination has been found.[37] There is relatively little known about how practitioners form TEAM diagnoses and weigh pieces of diagnostic information. If diagnosis is variable across different clinical conditions, we need to understand why and the implications for clinical practice.

It is difficult to make comparisons across studies of inter-rater reliability, since study design clearly influences interpretation of results. For example, some studies involving several practitioners have reported level of agreement on any syndrome for pairs of practitioners and then averaged them,[28,30–32] whereas others have reported agreement across all practitioners involved in the study on specific syndromes of a condition.[27] Obviously, the greater the number of practitioners involved, the less likely there is to be agreement. Many studies have not used formal statistical methods to assess reliability.

Factors that influence reliability

Several factors may potentially affect consistency of clinical observations, including practitioner variability due to differences in clinical education and experience, time of assessment (clinical signs and symptoms may change within hours), emphasis placed on different diagnostic techniques, and the inherent subjectivity of particular clinical observations.[30] TCM is diverse in its schools of thought and practices;[38,39] therefore, some disagreement would not be unexpected.

Other factors include differences in definitions of what they are assessing and differing notions of what they regard as "normal."[4] As King and colleagues found, when a standardized pulse-taking procedure was used with clear operational definitions, agreement was greater than 80% between 2 practitioners for 10 of 16 pulse categories.[18]

The statistical measure, that of the κ coefficient, is not without limitations. The κ coefficient varies with prevalence[37,40] and is influenced by the number of possible response categories.[37,41] Therefore, caution needs to be taken in making comparisons between studies.[37]

How questions are posed and studies are designed will determine, in part, results of inter-rater reliability studies. For example, forced choice from a predetermined list of patterns is problematic. It makes the *a priori* assumption that the patterns listed are definitive, that is, that there is agreement within the literature on these categories. This is not always the case. Few studies have attempted to empirically investigate the relative frequencies of TCM patterns in particular diseases. Although it increases the likelihood of agreement, choosing from a list of possible diagnoses does not mirror real-life practice.

Table 5. Inter-rater Reliability Studies of Pattern Diagnosis

Authors	*Subjects*	*Observers*	*Design features*	*Statistical test*	*Results*
Ogawa 1978[15]	5 Subjects	4 Meridian therapy practitioners	Reliability of pattern diagnosis using "four inspections" in Japanese meridian therapy. Observers examined subjects independently and were required to choose between four possible *sho* diagnoses.	No formal statistical analysis.	In 3 of the 5 subjects, 3 of the 4 observers agreed on *sho*. Ogawa argued that the levels of agreement were better in this study where various diagnostic data are integrated to select the pattern than when individual diagnostic techniques are used.
Birch 1997[a]	9 Subjects	5 Meridian therapy practitioners	Reliability of pattern diagnosis in Japanese meridian therapy based on pulse diagnosis, abdominal diagnosis, and questioning.	Kendall's coefficient of concordance test.	Significant correlation ($W = 0.29$, $\chi^2 = 10.39$, D.F. $= 4$, $p = 0.03$).
Birch 1999[b]	43 Subjects	2 Meridian therapy practitioners	Follow-up from above study.	Kendall's coefficient of concordance test.	Significant correlation ($W = 0.18$, $\chi^2 = 6.00$, D.F. $= 1$, $p = 0.01$).
Coeytaux et al. 2006[23]	37 Patients with frequent headaches	3 TCM practitioners	Observers examined participants individually. Observers required to choose from fixed diagnoses of *Qi*, Blood, *Yang*, *Yin*, Phlegm, Wetness, or "Other" disharmony and to choose the type of dysfunction (Rising, Stagnation, Deficiency, "Other"). Practitioners also required to choose which organ/meridian was involved from a fixed list (LV, GB, Sp, St, SI, LI, UB, Kid, Other) and name up to 8 acupoints.	No formal statistical tests used.	65% of participants were diagnosed at having *Qi* disharmony by at least 2 practitioners. Little agreement among practitioners for diagnoses of *Yang*, Blood, *Yin*, or Phlegm disharmony that were identified in at least half of the participants. Liver 3 (*Tai Chong*), Large Intestine 4 (*He Gu*), Governing Vessel 20 (*Bai Hui*) were acupoints commonly chosen by the 3 TCM acupuncturists.
Dang & Zaslawski 2003[17]	One subject with cystic fibrosis	6 TCM practitioners (4 experienced practitioners and 2 recent graduates)	Practitioners asked to diagnose, describe pulse and tongue, and to suggest a treatment principle and two acupuncture point prescriptions.	Not reported.	Low level of reliability for pulse and tongue diagnosis. Greater level of agreement concerning diagnostic conclusion.
Hogeboom et al. 2001[24]	6 Patients with low-back pain	6 TCM practitioners	Practitioners required to choose from 11 diagnostic categories that were specified by the investigators. Practitioners could write in additional diagnoses and make multiple diagnoses.Practitioners were also asked to propose an acupuncture treatment.	κ Coefficient reported as used; however, κ coefficient values not reported.	Median number of diagnoses made across 6 practitioners was 3.Reasonably high level of agreement on broad diagnostic categories (Kidney deficiency [any type] and Stagnation [any type]) but poor agreement on the 11 more specific syndromes.

					Substantial variation in treatment recommendations found with 65 different acupuncture points recommended at least once. Very little overlap in acupoints recommended by different practitioners for same patient.
Kalauokalani et al. 2001[25]	1 Patient with chronic low-back pain	7 TCM practitioners	Study used practitioners trained in TCM school and a screening questionnaire utilized to select acupuncturists whose approaches were similar. Practitioners required to record diagnosis and provide a first treatment including specific acupuncture points, duration of insertion, and whether additional electro-acupuncture stimulation or heat should be used.	No formal statistical analysis used.	A total of 6 diagnostic subcategories were assigned by practitioners. All practitioners assigned 2 or 3 diagnoses. *Qi* stagnation was the most commonly assigned (6/7 practitioners) followed by Blood stagnation (5/7 practitioners) with no other diagnoses recorded by more than 2 practitioners. Much variation found in acupuncture points selected by practitioners who recommended between 5 and 14 acupoints (requiring 7–26 needles, taking into account bilateral needling of some points). Of the 28 acupoints chosen by the 6 practitioners, only 14% were prescribed by 2 or more practitioners.
MacPherson et al. 2003[26]	148 Patients with low-back pain	5 TCM practitioners	A short list of syndromes was developed by the researchers and practitioners. Practitioners required to choose from 3 predefined syndromes of low-back pain specified: *Qi* and blood stagnation, Bi Syndrome, Kidney Vacuity, plus a 4th diagnostic category "Other" for primary and secondary diagnoses. Study compared the diagnosis of one practitioner with five other practitioners for 87 patients.	κ coefficient.	Practitioners identified 2 or more syndromes for 65% of subjects (mean 1.7 per patient). Level of agreement for the 5 pairs of practitioners ranged from 47% (κ 0.02) and 80% (κ 0.67) for either primary or secondary diagnoses.
O'Brien et al. 2008[27]	45 Patients with hypercholes-terolemia	3 TCM practitioners	Practitioners required to write their own CM syndrome diagnoses and were not prompted or constrained on number of syndromes that could be written. Level of agreement assessed for 6 most common syndromes identified.	κ coefficient not applicable since number of categories open-ended.	More than 15 different single CM syndromes diagnosed. Level of agreement between 3 practitioners: 24% for 1 syndrome (Spleen *Qi* deficiency) but poor on 5 other syndromes. When level of agreement between *at least two practitioners* was considered, level of agreement on the diagnosis of Spleen *Qi* deficiency increased to 86%.

(continued)

Table 5. (Continued)

Authors	*Subjects*	*Observers*	*Design features*	*Statistical test*	*Results*
Sung et al. 2004[28]	39 Patients with irritable bowel syndrome (IBS) in phase 1. In phase 3: 15 IBS and non-IBS patients	4 practitioners (six pairs)	Phase 1 of study: practitioners required to choose from 4 possible syndrome diagnoses (Stagnation of liver *Qi* and vacuity of the spleen; vacuity of spleen-Q*i*; accumulation of Damp-Heat in the spleen and stomach and vacuity of spleen and kidney). Practitioners also asked to choose 1 of 4 possible corresponding treatment principles and herbal medicinal formulae. Following a second phase of the study in which the 4 practitioners were asked to see 15 IBS patients together and discuss diagnosis and reach consensus, a 3rd phase of the study was conducted in which practitioners were required to choose a diagnosis in IBS and non-IBS patients.	κ coefficient.	Phase 1: Mean percentage agreement and mean κ's (average over the 6 pairs) in diagnosis and treatment principle were 57% (κ 0.11) and 58% (κ 0.16), respectively. Compared with phase 1, there were statistically significant improvements ($p = 0.002$) in agreement rate of the 4 practitioners: mean level of agreement in diagnosis and treatment principle improving to 80% (κ 0.34) and 81% (κ 0.37), respectively. Reliability of treatment regime improved from phase 1 to phase 3: mean level of agreement for 6 pairs of practitioners increased from 52% (κ 0.29) to 80% (κ 0.34).
Zhang et al. 2004[30]	39 Patients with rheumatoid arthritis (RA)	3 TCM practitioners	Pilot study conducted on 19 RA patients in order to compile list of syndromes, resulting in 15 categories that were condensed to 10 categories. Practitioners required to choose from a list of patterns and their corresponding treatment principles with an option to write other diagnoses not included in the list. Patients examined separately by each practitioner.	κ coefficient.	Mean level of agreement across 3 pairs of practitioners on any TCM syndrome diagnosis was calculated. Average consensus among 3 pairs of practitioners on TCM syndrome diagnosis: 28.2% (κ 0.26), ranging from 25.6% (κ 0.23) to 33.3% (κ 0.30). When a definition of partial agreement* on syndrome diagnosis was used, mean level of agreement between three pairs of practitioners increased to 64.8%, ranging from 48.7% to 84.7% (κ 0.55–0.87). Mean level of agreement on herbal medicine prescription across 3 pairs of practitioners was 28.2% (mean κ 0.28). When partial agreement definitions were used, the average level of agreement for 3 pairs of practitioners increased to 63.2 % (range 56.4–76.9%, κ 0.55–0.80). The degree to which herbal prescriptions agreed with textbook recommendations for particular patterns was found to be 93.2%. * Partial agreement on diagnosis defined as "two diagnoses that contain a mutual inclusive pattern with one diagnosis having an additional pattern" (Zhang et al. 2004, p. 59).

Zhang et al. 2005[31]	40 Patients with rheumatoid arthritis	3 TCM practitioners	Comparison study to previous study (Zhang et al. 2004). Practitioners examined patients separately.	κ coefficient.	Mean level of agreement for TCM diagnoses 31.7%. No statistically significant difference between result of this study and previous study (Zhang et al. 2004). The degree to which herbal prescriptions agreed with textbook recommendations was 91.7% (range 85–100%).
Zhang et al. 2008[32]	42 Patients with rheumatoid arthritis	3 TCM practitioners	Similar study design as previous 2 studies except in this study, a prior training session of TCM diagnostic procedures was conducted with an open case discussion. Same TCM practitioners utilized in this study as in previous study (Zhang et al. 2005).	κ coefficient.	Level of agreement for 3 pairs of practitioners ranged from 64.3% (κ 0.49) to 85.7% (κ 0.76). Significant improvement in level of agreement on diagnosis in this study compared to 2 previous studies (Zhang et al. 2004, Zhang et al. 2005).
Zell et al. 2000[29]	23 Post-menopausal women with hot flushes	9 TCM practitioners	Practitioners required to select from more than 30 different syndromes. Practitioners could also write their own diagnoses.	No formal statistical analysis.	High level of agreement on a diagnosis of Kidney *Yin* vacuity (Kidney *Yin* deficiency present yes or no). Kidney *Yin* deficiency diagnosis being made by 7/9 practitioners for 16/23 women and at least 5/9 practitioners in all 23 women. Other diagnoses made much less often and with a higher level of disagreement than agreement between practitioners.

[a]Birch S. An exploration with proposed solutions of the problems and issues in conducting clinical research in acupuncture [Ph.D. thesis]. Exeter University, 1997.
[b]Birch S. Preliminary investigations of inter-rater reliability of traditional based acupuncture diagnostic assessments [unpublished manuscript]. 1999.
TCM, Traditional Chinese Medicine; W, Kendall's coefficient of concordance; D.F., degrees of freedom.

Limitations of study designs need to be taken into account in interpreting results. For example, MacPherson and colleagues' study on low-back pain used predefined syndromes and compared the diagnoses of 1 practitioner with those of 5 others.[26] In this study, limitations included the fact that the level of agreement measured depended crucially on the 1 practitioner whose diagnoses were compared with those of the other 5.[26]

Another problem with attempts to establish the reliability of TCM pattern diagnoses is that it presupposes that they exist. This may not be the case. As pointed out by O'Brien and colleagues,[27] the applicability of *bian zheng lun zhi* (pattern differentiation and treatment determination) to biomedically defined clinical entities (for example, human immunodeficiency virus/acquired immune deficiency syndrome and hypercholesterolemia, which have no historical precedence of treatment), has not been established. Scheid[39] points out that the emphasis on *bian zheng lun zhi* as a *central* and *defining* tenet of TCM is recent, born around the 1950s, and that for long periods in Chinese medicine history, diseases rather than patterns were emphasized as diagnostic classifiers. Therefore, we need to be careful in making assumptions about the applicability of TCM pattern diagnoses to more "contemporary" and/or biomedically defined diseases or clinical entities that do not have a substantial history of treatment with TEAM.

Implications of understanding reliability

TEAM practice has long rested on the weight of experiential evidence built on thousands of years of systematic observation and documentation.[42] Increasingly, TEAM systems are being studied using scientific research methods. This is likely to be part of a cultural shift that has occurred in Western medicine, one that has shifted value from peer opinion to scientific research as the basis of medical practice and clinical decision-making.[1,43,44] Although it is tempting to dismiss the growing interest in scientific research into TEAM as simply an (unnecessary) attempt to justify TEAM practices to Western medicine, this would be erroneous. Although there are inherent difficulties in applying research methods to TEAM due to fundamental differences in paradigms underpinning TEAM compared with biomedicine, scientific research methods are simply tools that, when judiciously used, may be used to scrutinize the theories and practices and further develop them.

Understanding the reliability of TEAM diagnostic processes has implications for research and education. For example, if aspects of TEAM diagnoses are more unreliable, teachers and practitioners need to be aware of this, and ways need to be developed to improve reliability. King and colleagues[18] provide an example of this in their development of a standardized pulse-taking procedure and precise operational definitions that resulted in greater than 80% agreement between 2 practitioners on 10 of 16 pulse categories. Relatively little is known about the weight that practitioners place on various pieces of diagnostic information. However, if practitioners are placing emphasis on diagnostic data that are unreliable, there may be less confidence that the eventual diagnosis and treatment are optimal.

There are implications for clinical research. For example, if an herbal medicine study is designed to test efficacy of an herbal formula in a particular diagnostic subpopulation (i.e., that have a particular pattern), unless pattern diagnosis is reliable, there can be less confidence that the right study participants will be included. In studies that investigate efficacy of individualized (pattern-specific) treatment, if TCM pattern diagnosis is not reliable, how can one interpret whether correct treatment was given and therefore whether the treatment was efficacious or not? If TCM patterns are to be included in clinical studies, their reliability does need to be established.

Ways forward

If researchers are to attempt to establish reliability of TEAM diagnoses, a suggested approach is to first conduct a thorough literature search to ascertain what patterns are recorded and to establish whether the sources are dependable.[44,45] Appearance in a textbook is no guarantee of authenticity of information. A more thorough step would be to ascertain empirically whether distinct patterns in a particular disease or clinical entity do in fact occur. A means by which this could be achieved would be to conduct a TEAM examination in a study population with a particular disease using, as a control, a study population without the disease.

There are ways in which reliability of pattern diagnosis can be enhanced. As suggested by MacPherson and colleagues,[26] if treatment is individualized to the TEAM pattern, pre-trial training could be offered to all participating practitioners, thereby ensuring a minimum level of reliability in pattern diagnosis. They cite a study on acupuncture treatment of depression[46] as an example. Several groups have reported that after additional training to improve consistency of technique and agreement between practitioners, reliability increases from initial evaluations to post-training evaluations.[18,28,32] A final method can be to have 2 or more clinicians examine every patient in a clinical trial and to proceed with the treatment only after they have reached consensus on the diagnosis and treatment. The practitioners involved can also be pre-trained together in order to increase agreement before the study starts.

Conclusions

The investigation of reliability of TEAM diagnoses and treatments is in its formative stages. Reliability of pattern diagnosis has been found to be variable across types of practice and a range of diseases. Since diagnosis guides treatment, this has implications for education and clinical studies. If the observation and analysis of diagnostic data and patterns of diagnosis are not found to be reliable, it is necessary to understand why and develop strategies to improve it. We have seen that in particular studies mentioned earlier, development of clear definitions and training can help reliability. Until diagnostic data collection and pattern diagnosis are shown to be reliable, there can be little justification for inclusion in clinical studies of TEAM. Therefore, it is important that researchers develop strategies to improve reliability.

Disclosure Statement

The authors declare that they have no competing interests.

References

1. Birch S, Felt R. Understanding Acupuncture. Edinburgh: Churchill Livingstone, 1999.
2. Murakami M. Lecture given in Newton, Massachusetts, October 1998.
3. Ogawa T. Workshop given at the New England School of Acupuncture, Watertown, MA, April 1996.
4. Abramson JH. Surveys Methods in Community Medicine. New York: Churchill Livingstone, 1990.
5. Landis RJ, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.
6. Anthony H. Clinical research questions to ask and the benefits of asking them. Complem Med Res 1989:3:3–5.
7. Birch S. Naming the un-nameable: A historical study of radial pulse six position diagnosis. J Acu Soc New York 1994;1:19–32.
8. Birch S. Testing the claims of traditionally based acupuncture. Complement Ther Med 1997;5:147–151.
9. Bowling A. Choice of health indicators: The problem of measuring outcome. Complement Med Res 1988;2:43–63.
10. Canter D. A research agenda for holistic therapy. Complement Med Res 1987;2:104–116.
11. Vincent CA. Acupuncture as a treatment for chronic pain. In: Lewith GT, Aldridge D, eds. Clinical Research Methodology for Complementary Therapies. London: Hodder and Stoughton, 1993:289–308.
12. Debata A. Experimental study on pulse diagnosis of rokubujoi. Jpn Acup Mox J 1968;17:9–12.
13. Kurosu Y. Experimental study on the pulse diagnosis of rokubujoi 11. Jpn Acup Moxib J 1969;18,3:26–30.
14. Matsumoto T. Experimental study on fukushin (abdominal palpation). Jpn Acup Mox J 1968;17:13–16.
15. Ogawa T. To establish new "Chinese medicine": Searching for the contemporary significance of the "meridian controversy" [in Japanese]. Chin Med 1978;1:151–158.
16. Kass R. Traditional Chinese Medicine in San Francisco: Reliability of traditional Chinese pulse diagnosis. Social Welfare Department, University of California, Berkeley, 1987.
17. Dang T, Zaslawski CJ. A pilot study of the diagnostic reasoning processes used by practitioners of Traditional Chinese Medicine. The Fourth Australasian Acupuncture and Chinese Medicine Conference, Victoria University of Technology, Melbourne, Australia, 1998. In: Zaslawski C. Clinical Reasoning in Traditional Chinese Medicine: Implications for Clinical Research. Clin Acupunct Oriental Med 2003;4:94–101.
18. King E, Cobbin D, Walsh S, Ryan D. The reliable measurement of radial pulse characteristics. Acupuncture Med 2002;20:150–159.
19. King E, Cobbin D, Ryan D. The reliable measurement of radial pulse: Gender differences in pulse profiles. Acupuncture Med 2002;20:160–167.
20. Walsh S, Cobbin D, Bateman K, Zaslawski C. Feeling the pulse: Trial to assess agreement level among TCM students when identifying basic pulse characteristics. Eur J Oriental Med 2001;3:25–31.
21. O'Brien KA, Abbas E, Zhang J, et al. Understanding the reliability of diagnostic variables in a Chinese medicine examination. J Altern Complement Med 2009;in press.
22. Kim M, Cobbin D, Zaslawski C. Traditional Chinese medicine tongue inspection: A examination of the inter- and intrapractitioner reliability for specific tongue characteristics. J Altern Complement Med 2008;14:527–536.
23. Coeytaux RR, Chen W, Lindemuth CE, Tan Y, Reilly AC. Variability in the diagnosis and point selection for persons with frequent headache by traditional Chinese acupuncture. J Altern Complement Med 2006;12:863–872.
24. Hogeboom CJ, Sherman KJ, Cherkin DC. Variation in diagnosis and treatment of chronic low back pain by traditional Chinese medicine acupuncturists. Complement Ther Med 2001;23:153–155.
25. Kalauokalani D, Sherman KJ, Cherkin DC. Acupuncture for chronic lower back pain: Diagnosis and treatment patterns among acupuncturists evaluating the same patient. Southern Med J 2001;94:486–492.
26. MacPherson H, Thorpe L, Thomas K, Campbell M. Acupuncture for lower back pain: Diagnosis and treatment of 148 patients in a clinical trial. Complement Ther Med 2003;12:38–44.
27. O'Brien KA, Abbas E, Zhang J, et al. An investigation of the reliability of Chinese medicine diagnosis according to eight guiding principles and zang-fu theory in hypercholesterolaemic Australians. J Altern Complement Med 2009; 15:259–266.
28. Sung JJY, Leung WK, Ching JYL, et al. Agreements among traditional Chinese medicine practitioners in the diagnosis and treatment of irritable bowel syndrome. Aliment Pharmacol Ther 2004;20:1205–1210.
29. Zell B, Hirata J, Marcus A, et al. Diagnosis of symptomatic postmenopausal women by traditional Chinese medicine practitioners. Menopause 2000;7:129–134.
30. Zhang GG, Bausell B, Lao L, et al. The variability of TCM pattern diagnosis and herbal prescription on rheumatoid arthritis patients. Altern Ther Health Med 2004;10:58–63.
31. Zhang G, Lee WL, Bausell B, et al. Variability in the traditional Chinese medicine (TCM) diagnoses and herbal prescriptions provided by three TCM practitioners for 40 patients with rheumatoid arthritis. J Altern Complement Med 2005;11:415–421.
32. Zhang GG, Singh B, Lee WL, et al. Improvement of agreement in TCM diagnosis among TCM practitioners for persons with the conventional diagnosis of rheumatoid arthritis: Effect of training. J Altern Complement Med 2008;14:381–386.
33. Broffman M, McCulloch M. Instrument assisted pulse evaluation in the acupuncture practice. Am J Acup 1987;14:255–259.
34. Nicole G. A pilot study in which five fourth year students diagnose the pulse of eight patients to indicate the apparent effectiveness of its fourth year students in taking the pulses and its own apparent effectiveness in the teaching of pulse diagnosis. [Undergraduate Independent Research Project]. Sydney: Acupuncture Colleges (Australia) 1990. In: Zawslawski C. Clinical Reasoning in Traditional Chinese Medicine: Implications for Clinical Research. Clin Acup Oriental Med 2003;4:94–101.
35. Wiseman N, Ye F. Translation of Chinese medical pulse terms, taking account of the historical dimension. Class Acup Orient Med 1999;1:55–60.
36. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? JAMA 1994;271:389–391.
37. Joshua AM, Celermajer DS, Stockler MR. Beauty is in the eye of the examiner: Reaching agreement about physical signs and their value. Intern Med J 2005;35:178–187.

38. Scheid V. Not very traditional, nor exactly Chinese, so what kind of medicine is it? TCM's discourse on menopause and its implications for practice, teaching and research. J Chin Med 2006;82:5–20.
39. Scheid V. Chinese Medicine in Contemporary China. Plurality and Synthesis. Durham and London: Duke University Press, 2002.
40. Hansen M, Sindrup SH, Christensen PB, et al. Interobserver variation in the evaluation of neurological signs: observer dependent factors. Acta Neurol Scand 1994;90:145–149.
41. Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall, 1996.
42. Birch S, Lewith G. Acupuncture research, the story so far. In: MacPherson H, Hammerschlag R, Lewith G, Schnyer R, eds. Acupuncture Research: Strategies for Building an Evidence Base. London: Elsevier, 2007:15–35.
43. Tonelli MR 1999. In defense of expert opinion. Acad Med 1999;74:1187–1192.
44. MacPherson H, Sherman K, Hammerschlag R, et al. The clinical evaluation of traditional East Asian systems of medicine. Clin Acup Oriental Med 2002;3:16–19.
45. Birch S. Testing the claims of traditionally based acupuncture. Complement Ther Med 1997;5:147–151.
46. Allen JJB, Schnyer RN, Hitt SK. The efficacy of acupuncture in the treatment of depression in women. Psychol Sci 1998;9:397–401.

Address reprint requests to:
Kylie A. O'Brien, Ph.D.
Faculty of Health, Engineering and Science
Victoria University
McKechnie Street
St. Albans, Victoria 3021 Melbourne
Australia

E-mail: kylie.obrien@vu.edu.au