REVIEW ARTICLE

Reliability studies of diagnostic methods in Indian traditional Ayurveda medicine: An overview

Vrinda Hitendra Kurande, Rasmus Waagepetersen[1], Egon Toft[2], Ramjee Prasad[3]

Clinical Science, Departments of Health Science and Technology, [1]Departments of Mathematical Sciences, [2]Departments of Health Science and Technology, [3]Center for TeleInFrastructure, Aalborg University, Denmark

ABSTRACT

Recently, a need to develop supportive new scientific evidence for contemporary Ayurveda has emerged. One of the research objectives is an assessment of the reliability of diagnoses and treatment. Reliability is a quantitative measure of consistency. It is a crucial issue in classification (such as *prakriti* classification), method development (pulse diagnosis), quality assurance for diagnosis and treatment and in the conduct of clinical studies. Several reliability studies are conducted in western medicine. The investigation of the reliability of traditional Chinese, Japanese and Sasang medicine diagnoses is in the formative stage. However, reliability studies in Ayurveda are in the preliminary stage. In this paper, examples are provided to illustrate relevant concepts of reliability studies of diagnostic methods and their implication in practice, education, and training. An introduction to reliability estimates and different study designs and statistical analysis is given for future studies in Ayurveda.

Key words: Ayurveda, diagnostic methods, kappa statistics, reliability, traditional medicine

INTRODUCTION

Ayurveda as a traditional and holistic medicine has a sound philosophical and experiential basis.[1] Long historical use has been seen as documentation of the efficacy; however, there is lack of quantitative studies of concepts such as reliability as evaluated in the modern medicine. The term reliability means "consistency" of a measure.[2] Diagnosis is considered reliable if it gives consistent results under similar conditions. Assessment of reliability is essential because consistent diagnosis leads to consistent treatment.

In research, there are eight criteria for evaluating the patient-based outcome measures for any specific clinical trial: Appropriateness, reliability, validity, responsiveness, precision, interpretability, acceptability, and feasibility.[3] In Ayurveda, outcome measures may be the findings of physical examinations or the scores on questionnaires collecting information, for example, on body constitution (*prakriti*), life style, diet, and quality of life. If these outcome measures or diagnostic variables such as bodily humors/energies (*dosha*), tissues (*dhatu*), metabolic waste (*mala*), metabolic fire (*agni*), subtle life force (*prana*), material life force (*ojas*), toxicity (*ama*), body parts (*avayava*), vital body part (*marma*), 5 elements (*mahabhoota*) are to be included in the research protocol, the prerequisite is that they should be valid and reliable.[4]

In Ayurveda, diagnostic methods (such as pulse diagnosis) often rely on some degree of subjective interpretation by physicians. If the physicians cannot agree on the interpretation, the results will be of little use. Hence, reliability studies are necessary for quality assurance in the conduct of clinical studies and practice.[5]

In this paper, we provide a selected summary of the reliability studies of the physical examination commonly used in Ayurveda and different traditional medicines from Asia. It is noteworthy to reflect on the reliability studies carried out in traditional Chinese medicine (TCM), Japanese traditional medicine: *Toyohari* meridian therapy (TMT), and Sasang Constitutional Medicine (SCM). These traditional medicines have some similarities with Ayurveda.[1,6] One

Address for correspondence:
Dr. Kurande Vrinda Hitendra, Aalborg University,
Frederik Bajers Vej 7, D-2/105, 9220 Aalborg Øst, Denmark.
E-mail: v_kurande@yahoo.com

Received: 19-Feb-2013
Revised: 21-Mar-2013
Accepted: 15-Apr-2013

Access this article online	
Quick Response Code:	Website: www.jaim.in
	DOI: 10.4103/0975-9476.113867

common finding is that diagnostic methods of all traditional medicines rely more on the physicians reading of the patient's signs and symptoms than on laboratory findings.

The scope of this paper is limited to inter and intra rater reliability for specific diagnostic methods: Pulse (*nadi*) diagnosis, body constitution diagnosis and tongue diagnosis. In final, future perspectives of reliability studies in Ayurveda are discussed.

METHODS

A literature review is conducted using electronic databases "PubMed" "Google Scholar" and "Scopus." The review was conducted with an interactive strategy of combining the keywords "reliability," "agreement," "traditional medicine," "alternative medicine," "Ayurveda" "complementary medicine," "Chinese medicine," "Sasang medicine," "*Toyohari* medicine." Further, advanced or refined search was carried out using the key words "diagnostic methods," "physical examination," "pulse diagnosis," "body constitution," "*prakriti,*" and "tongue diagnosis." Furthermore, reference lists from previous systematic reviews were browsed.[7-9] Articles were limited to those in the English language.

OBJECTIVES

The objective of this study was to provide information about how the reliability studies can be designed and conducted for Ayurvedic diagnostic methods. Importance of the reliability studies in practice and clinical trials will be discussed in relation to illustrative case studies.

WHAT IS RELIABILITY?

Reliability denotes consistency of a measure. Reliability provides information about the amount of error inherent in any diagnosis score or measurement, where the amount of measurement error determines the validity of the study results or scores.[5]

Reliability verses validity

Validity and reliability concepts can easily be misunderstood. Validity is analogs to accuracy. A test/instrument is valid when it measures, what it is intended to measure. The test is reliable when it produces same results under identical conditions. Thus, reliability does not denote validity [Figure 1]. For example, if a person, who weighs 50 kg steps on a weighing scale 4 times and gets readings of 45, 48, 40, and 54 kg the scale, is not reliable and if it consistently reads "45 kg" it is reliable, but not valid. If it reads "50 kg" each time, it is reliable and valid. A test that is not reliable cannot be completely valid. Measures of validity of diagnostic procedures are commonly quantifying the ability of the procedures to distinguish individuals with and without a certain disease. Basic measures for this purpose, such as sensitivity and specificity, likelihood ratios, positive, and negative predictive values are described elsewhere.[10,11] More elaborate measures of validity for, e.g., psychological testing are presented in[12] It is essential that a diagnosis is reliable and valid. However, in Ayurveda, the problem with assessing validity is that there is lack of "gold standard" to compare with. E.g., for pulse diagnosis the diagnosis can only be obtained from a doctor's judgments. However, since different doctors may obtain different diagnoses, we do not know which one is the true diagnosis that all other diagnoses should be compared with.

TYPES OF RELIABILITY

There are several types of reliability estimates. [2,13] The terms "reliability" and "agreement" are often used interchangeably.[14] The two concepts are conceptually distinct. Reliability parameters are the most appropriate when the research questions concerns with the distinction of persons. However, parameters of the agreement are preferred when the aim of the study is to measure change in health status, which is often the case in clinical practice. However, similar study designs are used for examining these two concepts. Guidelines are also available for reporting reliability and agreement studies.[15]

Intra-rater reliability

It is also known as "test-retest reliability" or "repeatability." This type of reliability is used to assess the consistency of a test across time. Intra-rater reliability testing is the process by which a measurement tool or method can be shown to give similar results when used by same raters at different time for the same group of subjects. Examples are given in Tables 1 and 2 [Figure 2].

Inter-rater reliability

Inter-rater reliability or "reproducibility" denotes in clinical settings the extent to which doctors agree with each other in their diagnosis and treatment. This is assessed by allowing two or more independent doctors carry out independent assessments of the same patient. The scores are then compared to determine the consistency of the doctor's estimates. Examples are given in Tables 1 and 2 [Figure 3].

Inter-method reliability

Inter method reliability is assessed by comparing two different methods or tests [Figure 4]. When a new method is proposed its value can be assessed only by comparing with another established technique (gold standard),

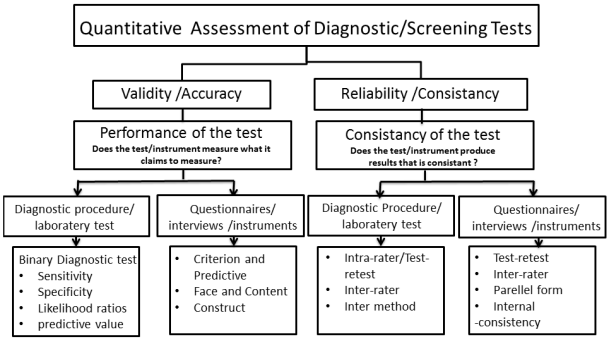

Figure 1: Quantitative assessment of diagnostic methods/tests

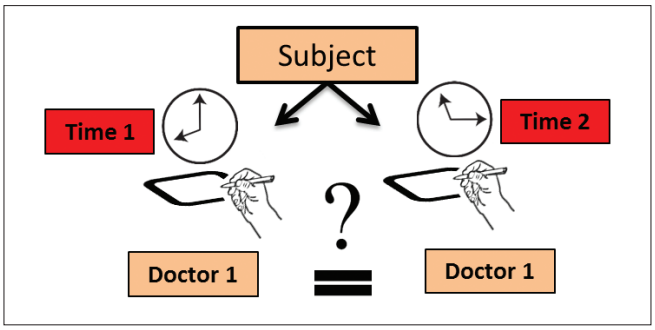

Figure 2: Intra-rater reliability

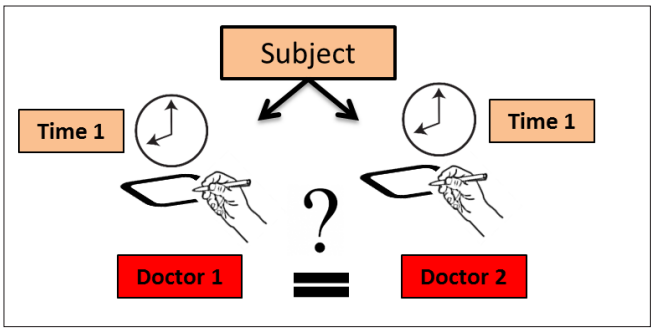

Figure 3: Inter-rater reliability

rather than with the true quantity being measured. It is not possible to verify that either method gives definitely correct measurement, so it is necessary to assess the degree of agreement between them. For example, an automated oscillometric blood pressure monitor was compared with the auscultator mercury sphygmomanometer.[16] This comparison between the different methods is assessed in a similar manner as for intra-rater reliability. Analysis on the agreement between two methods of clinical measurement is proposed by Bland and Altman.[17]

In brief, inter and intra rater reliability estimates are used when the raters (doctors) are part of the experiments. To establish a new technique or method, inter method reliability is assessed. While dealing with forms or questionnaires/instruments, inter method reliability is termed parallel-forms reliability. A way of discovering which questions are more informative is to use two questionnaires in parallel and finalize one that is reliable. Further, reliability of each item (questions) is estimated by internal consistency reliability as explained in the next section.

Internal consistency reliability

This form of reliability is used to assess the consistency of results across items on the same test or questionnaire [Figure 5]. This can be determined by average inter-item correlation, average item total correlation, and split-half correlation.[2,13] The internal consistency of Sasangin diagnosis questionnaire analysis was carried out using the data collected from 423 respondents. The questionnaire consisted of a total of 229 items. Cronbach's alpha coefficient (above 0.50) showed that all the categories can be accepted as being reliable scales, meaning this questionnaire is acceptable for surveying purposes.[18]

HOW TO CONDUCT RELIABILITY STUDIES IN AYURVEDA

Intra-rater reliability

Intra-rater reliability can be conducted by a single or more raters on same subjects. However, reliability estimates may vary according to the time interval between repeated measures. Assessment of intra-rater reliability is difficult for some direct observable signs and symptoms, since results may be influenced by the observer's memory or attempts at consistency in observations.[19] Hence, it is necessary to keep adequate time interval in between diagnoses to avoid the carryover effect of first diagnosis. In intra-rater assessment of Sasang constitution (SC), six SCM experts diagnosed SC of 86 participants twice independently with 1 year between first and second diagnosis. The reliability was moderate.[20] The time interval of 1 year was necessary to avoid the carryover effect of the first diagnosis.

On the other hand, pulse and tongue characteristics may change within hours or a day. It makes assessment of intra-rater reliability more difficult. It is possible only if such studies are conducted in a short time to avoid possible variation in pulse or other symptoms. Furthermore, blinding and randomization is necessary to avoid carry over effect of the previous diagnosis. In an intra-rater assessment of pulse diagnosis, an Ayurvedic pulse diagnosis expert diagnosed *doshaja* type pulse of 17 healthy subjects twice in a random order. The trial was conducted on the same day and within a short time period. To avoid carry over effect of first diagnosis the doctor did pulse diagnosis without seeing the subjects. The reliability was moderate. Furthermore, raters should be blinded in the sense that they should not be aware of the number of participants and number of rounds.[21,22]

Table 1: Reliability studies in different traditional medicines from Asia

Authors	Diagnostic method	Type of reliability	Subject studied	Observers	Study design	Statistical test	Results
Jang, 2012.[20]	Sasang constitution	Intra-rater reliability	86 healthy subjects	Six experts	Experts interviewed subjects twice with an interval of 1 year	Cohen kappa	Range for individual expert from 0.38 to 0.76
Kim *et al.*, 2008.[23]	TCM tongue inspection	Intra-rater reliability	Ten realistic tongue slides	30 TCM practitioners	Same practitioner used two data sets to evaluate tongue characteristic on the same tongue	Descriptive statistics, predominantly percentage frequency agreement	Only 2 subjects achieved higher than 80% agreement level for all tongue slides on all questions, with the highest intra-rater reliability is 88% followed by 82%
Kim *et al.*, 2012.[41]	TCT	Intra-rater reliability	50 tongue photographs	24 reliable assessors were selected from 60 oriental medical doctors	Raters TCT judgments and DTIS-measured values was examined to ascertain the reliability DTIS measurements	Fleiss' k for over all agreement and Pearson's correlation	Moderate (κ=0.56), The level of correlation between TCT judgments and DTIS measurements was high (0.76, *P*<0.01)
King *et al.*, 2002.[40]	TCM, pulse diagnosis	Inter-rater reliability	66 subjects and in a replication collection 30 subjects	Two rater	Radial pulse measurement; initial data collection (66 subjects) and in a replication collection (30 subjects) completed two months later	Percentage agreement	Averaged 80%
O'Brien *et al.*, 2009.[24]	Pulse diagnosis, abdominal diagnosis and *sho* diagnosis in TMT	Inter-rater	Sixty-two (62) Australians (22 males, 40 females) aged 20-65 years	Two TMT practitioners	Raters independently conducted TMT examinations	Proportion of agreement	Level of agreement for pulse depth-57%, Pulse speed-61% and pulse strength-77% For abdominal diagnosis; involvement of the lung, kidney, spleen, and liver abdominal regions was 58%, 53%, 35%, and 10%, respectively, primary *sho*-48% and for secondary *sho*-44%
Zhang *et al.*, 2004.[51]	TCM diagnosis on patients with RA, the tongue and pulse and a herbal prescription	Inter-rater	39 patients with RA	Three licensed acupuncturists	Practitioners examined the same patients separately, following the traditional four diagnostic methods. Patients filled out questionnaires and physical examinations, including observations of the tongue and palpation of radial pulse, were conducted by the 3 practitioners	Kappa statistics	0.28 (0.25-0.33 with kappas ranging from 0.23 to 0.30) little agreement among the 3 practitioners with respect to the herbal formulas prescribed
Zhang *et al.*, 2005.[52]	TCM diagnosis on patients with RA, the tongue and pulse and a herbal prescription	Inter-rater	40 patients with RA	Other three licensed acupuncturists	Same as above		0.31 (range, 0. 27-0.35) 3 TCM practitioners were at the same low level as previously reported
Zhang *et al.*, 2008.[53]	TCM diagnosis on patients with RA, the tongue and pulse and a herbal prescription	Inter-rater	42 patients with RA	Three licensed acupuncturists same as in second study	Same as above but after the training, an open case discussion and "real time" practice		0.73 (0.64-0.85). Statistically significant differences were found between this study and the two previous studies (*P*<0.001)
Ryu *et al.*, 2010.[38]	Cold and Heat pathologic Pattern identification in TCM	The internal consistency test	63 patients (Group A) and 64 patients (Group B) 10 items for each type	Cold-Heat Pattern Questionnaire	Same questionnaire was completed by group "A" and Group "B"	Cronbach's α coefficients	0.579 for the 10 Cold items and 0.718 for the 10 Heat items

Contd...

Table 1: Contd...

Authors	Diagnostic method	Type of reliability	Subject studied	Observers	Study design	Statistical test	Results
Yoo *et al.*, 2007.[18]	SDQ for assessment of Sasang constitutional medicine	Intra-rater reliability	88 questionnaires	Self-report structured questionnaire	Questionnaire was administer twice with an interval of 2 weeks	Pearson's correlation coefficients	0.44-0.74
		The internal consistency test	Total 223 items			SDQ items had three choices, Cohen's kappa coefficient for 3×3 was used	40 items showing "a low degree of concordance (kappa values <0.4)," 1 item showing "a high degree of concordance," and the remainder (182 items) showing "a moderate degree of concordance"

TCM=Traditional chinese medicine, TCT=Tongue coating thickness, DTIS=Digital tongue imaging system, TMT=*Toyohari* meridian therapy, RA=Rheumatoid arthritis, SDQ=Sasangin diagnosis questionnaire

In case of tongue diagnosis or observable signs, uses of photographs or video recording can be used for the repeated diagnosis after an adequate time interval. In TCM, tongue slides were used for the intra-rater reliability assessment.[23]

Inter-rater reliability

Inter-rater reliability is conducted by more than one rater on the same group of subject. In Japanese medicine, two experts independently carried out pulse, abdominal and *sho* diagnosis (primary and secondary patterns of disharmony) of 62 healthy subjects. Reliability was moderate for pulse diagnosis than *sho* diagnosis and abdominal diagnosis shown in [Table 1]. Low level of reliability on *sho* diagnosis suggests room for improvement.[24]

Several factors may potentially affect the reliability of clinical observations: E.g., practitioner's and patient's variability. Following factors need to be considered.

Raters

There is diversity in the practice of Ayurveda and inherent subjectivity in the clinical observation. Practitioner variability due to the practitioner's experience, education, specialization, and reliance on different traditions should be considered while interpreting results.[22] Frequently, the larger number of practitioners involved, the less likely is agreement. Moreover, it is reasonable to keep the raters blinded in the sense that they should not be aware that their diagnosis will be judged with the other raters. This will ensure that their behavior is not altered because of awareness of being observed.[25] Diagnosis should be conducted independently by each rater without communication.

Subjects

Subject's variability may have influence on the result because clinical signs and symptoms may change within some time limit such as for pulse diagnosis. It is recommended to conduct the study on the same day and the same time. Only, *prakriti* diagnosis can be conducted at any time as it remains unchangeable for whole life.

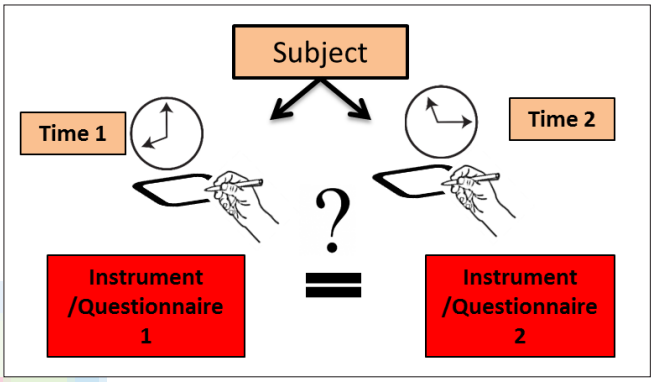

Figure 4: Inter-method reliability

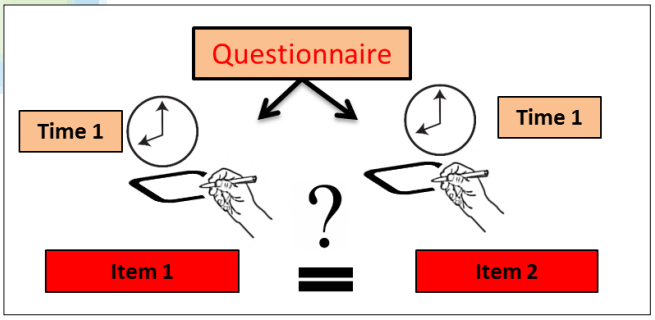

Figure 5: Internal consistency reliability

Diagnostic method

The degree of reliability is related to the properties of the methods or classifications that are used. Diagnostic methods exist in various forms, due to the different traditional practices. Different methods may have been adapted by different practitioners. Therefore, care should be taken in the study design that the same diagnostic method is used by all raters. Further, the specification and interpretation of diagnostic method is essential. When diagnostic methods are to be used in broader clinical contexts and daily practice,

Table 2: Reliability studies in Ayurveda

Authors	Ayurvedic diagnostic method	Type of reliability	Subject studied	Ayurvedic practitioners	Study design	Statistical test	Result
Rastogi, 2012[36]	PPAT	Inter-rater co-relation	26 healthy	Two	All the subjects examined on PPAT by both the raters independently	Correlation coefficient	Correlation coefficient for *kapha*-0.4074 (*P*<0.02), *pitta* 0.5245 (*P*-0.01), and *vata,* 0.8081 (*P*-0.001)
Prlic *et al.*, 2003[46]	Ayurvedic disease origin, diagnosis and treatment approach for inflammatory arthritis	Inter-rater reliability	three patients with inflammatory arthritis	Three	Ayurvedic practitioners checked subjects independently and asked to write ayurvedic disease origin, disease diagnosis, and treatment approach for each patient	No formal statistical analysis	Considerable agreement practitioners agreed upon 17 of 21 treatment groups
Dhruva *et al.*, 2012.[39]	*Prakriti* and *vikriti*	Inter-rater reliability	Three patients	13	Thirty minute videotaped recording of ayurvedic assessment including a history and limited physical exam was viewed to diagnose *prakriti* and *vikruti* and explain their rational for making a diagnosis	Cross sectional comparison and thematic analytic approaches were used to analyze qualitative data	Over all agreement level ranged 60-100%, *Prakriti*-mean 75%, *vikriti* mean 86%
Kurande *et al.*, 2012.[21]	*Doshaja* pulse diagnosis and body constitution diagnosis	Intra-rater reliability	17 healthy	One pulse diagnosis expert	A double-blinded, controlled, clinical trial expert diagnosed subject twice in a random order, body constitution was diagnosed by pulse and observation of hand (skin colour, hair, joint)	Weighted kappa, Landis and Koch scale	Pulse diagnosis-0.42 (Moderate agreement), Body constitution 0.65 (substantial agreement)
Kurande *et al.*, 2012.[22]	*Doshaja* pulse diagnosis	Inter and intra-rater reliability	20 healthy subjects 10 male and 10 female	15 ayurveda experts	A double-blinded, controlled, clinical trial. Practitioners diagnosed pulse twice blindly in a random order	Weighted kappa and Landis and Koch scale	Slight to moderate Intra-rater reliability (range k=0.10-0.55) and negative for two doctors. Low reliability

PPAT=Prototype *prakriti* analysis tool

reliability should also be investigated under conditions as close as possible to the clinical daily routine.[24]

Statistical analysis

The choice of statistical method depends on the type of data (nominal, ordinal, continuous), the sampling (at random, consecutive, convenience) and treatment of random and systemic error.[5] The intra-class correlation coefficient may be useful for measuring the reliability at continuous scale.[26] Moreover, proportion of a specific agreement,[27] reliability coefficient and graphical methods are also suggested for continuous data.[17] Cohen's kappa (K), Fleiss kappa, and weighted kappa statistics are indices of reliability for use with nominal scales.[28] Kappa statistics is a recognized measure of level of agreement beyond that expected by chance alone. It gives a quantitative measure of the magnitude of agreement between observers. Possible kappa values range from + 1 (perfect agreement) via zero (no agreement above that expected by chance) to - 1 (complete disagreement). An interpretation of kappa values in terms of level of agreement is given in Table 3.[29] Examples of kappa values and their interpretation for common clinical signs are provided in Table 4.[7] These examples provide evidence about the different levels of reliability of the physical examination and common clinical signs in western medicine.

In Ayurveda, most of the variables are nominal and categorical such as *dosha*, *prakriti*. Consequently, for Ayurveda it is relevant to understand kappa statistics. In weighted (K_w) kappa, disagreements of varying gravity (or agreements of varying degree) are weighted accordingly.[30] For example, the doctors would likely consider a diagnostic disagreement between "*vata*" and "*kapha*" to be more serious than between "*vata*" and "*vatakapha.*" If we use Cohen's kappa, it makes no distinction, implicitly treating all agreement (disagreement) equally. In an Ayurvedic pulse diagnosis study, additional interpretation of Cohen's weighted kappa statistic for analysis of categorical pulse and body constitution diagnosis was provided. For quantification of the reliability measure, weights were assigned based on the various compositions of *vata*, *pitta,* and *kapha*. A detailed presentation of weights based on a distance measure for pulse and body constitution diagnosis is presented in a reliabilities studies on pulse diagnosis.[21]

Comparisons of kappas across studies must be interpreted carefully because kappa values vary with prevalence.[28]

Moreover, the number of possible response categories of a test also influences kappa.[31] The kappa will be high when there are only two categories: E.g., presence or absence of a disease. As the number of categories increases, the kappa values will be smaller. Nevertheless, most of the medical literature on reliability has been reported in terms of kappa values and it remains a useful summary measure of reliability.

IMPLICATION OF RELIABILITY STUDIES

The results of the clinical trials conducted on many herbs and formulations could be improved by incorporating classical principles of Ayurveda diagnosis.[32] For this the prerequisite is that these variables should be reliable. In the following section, we will discuss body constitution, pulse, and tongue diagnosis studies from traditional medicine and their implication in research, education, and practice.

Body constitution diagnosis

Development of constitutional questionnaire

In Ayurveda, *prakriti* based prescription helps in enhancing the therapeutic effect as well as reduces the unwanted effects of the drug. For better results, it is important to include *prakriti* assessment in the clinical trial as inclusion/exclusion criteria.[33] There are few interesting studies indicating either a genetic or a biochemical basis for body constitutional types.[34,35] A pilot study on development and validation of a prototype *prakriti* analysis tool reported that *vata* and *pitta* constructs of *prakriti* identification have a significant inter-rater correlation ($P < 0.001$ and $P < 0.01$), whereas *kapha* has less ($P < 0.02$) correlation. It is concluded that *kapha* features are required to be designed more carefully to reach better consensus.[36] Some reliability studies on SC have been carried out in SCM.[20,37] Ayurvedic *prakriti* questionnaire includes three main categories that are *vata pitta* and *kapha* types. Internal reliability measures whether several questions that propose to measure the *vata*, *pitta,* and *kapha*, categories produce similar scores. Internal reliability is important while developing and evaluating a questionnaire. For example, a study on cold and heat pathologic pattern identification in TCM showed significant differences in the mean questionnaire scores between the cold and heat groups.[38] It is concluded that the questionnaire may be useful as an adjunct diagnostic tool.

In another study, the agreement between raters clinical rationale was assessed for Ayurveda diagnosis. Overall agreement on diagnoses ranged between 60% and 100% and was higher for *vikriti* (mean 86%) than *prakriti* (mean 75%) for thirteen participants.[39] Even when there was disagreement in diagnoses, the clinical rationale provided by physicians was consistent with the theoretical basis of Ayurvedic medicine. Interpretation of these studies is difficult because the study reported percentage of agreement. The result could be misleading because it does not take into account the agreement by chance.

Table 3: Interpretation of kappa values by Landis and Koch scale

Kappa value	Strength of reliability[29]
<0.0	Poor
0.01-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

Table 4: Comparisons of kappa values for common clinical signs

Sign[7]	Kappa value	Level of reliability
Heberden's nodes	0.78	Substantial
Size of goiter by examination	0.74	Substantial
Presence of goiter by inspection	0.65	Substantial
Abnormality of extra-ocular movements	0.77	Substantial
Forced expiratory time	0.70	Substantial
Signs of liver disease (e.g., jaundice, Dupytren's contracture, spider naevi)	0.65	Substantial
Dullness to percussion	0.52	Moderate
Wheezes	0.43-0.93	Moderate-perfect
Palpation of the posterior tibial pulse	0.60	Moderate
Palpation of dorsalis pedis pulse	0.51	Moderate
Hearing a systolic murmur	0.30-0.48	Fair-moderate
Chest expansion	0.38	Fair
Bronchial breath sounds	0.32	Fair
Crackles	0.30-0.63	Fair-substantial
Cough	0.29	Fair
Tachypnoea	0.25	Fair
Breath sound intensity	0.23-0.46	Fair-moderate

Pulse diagnosis

Development of a standardized pulse taking procedure and importance of training

Ayurvedic pulse diagnosis is the unique and non-invasive diagnostic method that determines the state of *dosha*; however, this is only justifiable if pulse diagnosis yields a consistent result. Many studies in TCM and TMT reported low to very good level of reliability for pulse diagnosis. An identified reason for the low reliability of TCM pulse diagnosis was the complex and ambiguously defined TCM pulse qualities and little systematic information.[8] So, effort was taken to develop a standardized pulse taking

procedure. A high level of reliability (80%) was observed when the study was conducted by developing concrete operational definitions for each of the characteristics of the pulse. However, the study reported percentage agreement instead of kappa value.[40] Thus, it is hard to judge the results.

In Ayurveda, in a double-blinded, controlled clinical trial, fifteen Ayurvedic doctors examined the *dosha* type of pulse of 20 (bio-medically defined) healthy subjects twice in a random order without seeing them. The doctors seem to favor different diagnoses since the proportions of ratings vary among doctors [Table 2].[22] Possibly, reliability of Ayurveda pulse diagnosis may be improved by standardizing the pulse taking procedure and by proper training.

Tongue diagnosis

Development of differential criteria on tongue diagnosis

A study on tongue diagnosis was conducted by 30 TCM practitioners. It is reported that the low level of reliability was due to inadequate operational definitions of both the tongue characteristics studied and of the inspection regions of the tongue.[23] When there is a need to develop standards for the diagnosis, it is obligatory that it should be carried out by reliable practitioners. For instance, 24 reliable raters (kappa value = 0.56) were selected from 60 rater to develop standards for judgment on the tongue coating thickness (TCT). The correlation between rater's TCT judgments and digital tongue imaging system judgments was high. Accordingly, it is proposed that thick coating of the tongue is that occupying approximately more than two-third of the tongue surface area.[41] Similarly, it will be beneficial to develop evidence based diagnostic guidelines for Ayurveda practice.

Finally, reliability studies are extended to disease diagnosis; specifically on lower back pain,[42] menopausal syndrome,[43] irritable bowel syndrome,[44] and headache[45] in TCM. Considerable agreement existed among three practitioners for the diagnosis and treatment of inflammatory polyarthritis despite Ayurvedic medicines individualized approach.[46]

FUTURE PERSPECTIVES

The diagnosis made during the Ayurvedic clinical evaluation of a patient should be consistent; Ayurvedic diagnostic variables are only justifiable if they are reproducible by different physicians for the same group of patients. Evidence of high reliability will improve the confidence among the doctors and these methods will possibly be incorporated into the clinical trials. If diagnosis is variable across different physicians, there is a need to understand the reason behind this variability. Moreover, to improve the quality and value of patient care, it is important to assess physician's performance in the clinical practice.[47] Based on the reliability results, clinical reliance should be given on reliable variables or methods.

Development of diagnostic guideline based on current scientific evidence is inevitable for contemporary Ayurveda. The study conducted by Patwardhan *et al.*[48] suggests that the Ayurvedic academicians are required to be trained in standard methods of research and the educational institutions should contribute in building up the evidence base for Ayurveda in the form of quality education and research as demonstrated in rheumatologic studies in Ayurveda.[49]

As observed in a few studies, if reliability is low, the key to improve the reliability is greater standardization of the most robust methods and a better understanding of the examination technique and its failings. Many studies show that training of professionals, improving the diagnostic instrument or method and a combination of both plays a significant role in greater reliability.[50] Furthermore, in TCM, a low level of reliability was observed in two subsequent studies on rheumatoid arthritis (κ value = 0.28 and κ value = 0.30 respectively).[51,52] Improvement in the level of reliability (κ value = 0.73) was observed after training sessions for the practitioners from study two [Table 1].[53]

Reliability can be improved by examining more frequently or examining same patient by more than one clinician. How to best increase the number of observations depends on the nature of the variation and diagnostic method. For example, variation in the hypertension diagnosis can be overcome by measuring blood pressure on several occasions by the same clinician.[54]

In conclusion, the reliability of diagnostic methods is of concern in research, education, and clinical practice. For contemporary Ayurveda, to be recognized as a credible health-care system, there is a need for rigorous reliability studies to be performed in the future.

REFERENCES

1. Patwardhan B, Warude D, Pushpangadan P, Bhatt N. Ayurveda and traditional Chinese medicine: A comparative overview. Evid Based Complement Alternat Med 2005;2:465-73.
2. Trochim, William M. The Research Methods Knowledge Base. 2nd ed. Available from: http://www.socialresearchmethods.net/kb/>(version current as of October 20, 2006) last accessed date : 13 February, 2013.
3. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials.

Health Technol Assess 1998;2:i-iv, 1-74.
4. Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use. 3rd ed. New York: Oxford University Press Inc.; 2003.
5. Dunn G. Statistical Evaluation of Measurement Errors: Design and Analysis of Reliability Studies. 2nd ed. London, UK: Arnold; 2004.
6. Lee SW, Jang ES, Lee J, Kim JY. Current researches on the methods of diagnosing sasang constitution: An overview. Evid Based Complement Alternat Med 2009;6:43-9.
7. 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;35:178-87.
8. O'Brien KA, Birch S. A review of the reliability of traditional East Asian medicine diagnoses. J Altern Complement Med 2009;15:353-66.
9. Zaslawski C. Clinical reasoning in traditional Chinese medicine: Implications for clinical research. Clin Acupunct Orient Med 2003;4:94-101.
10. Brenner H. Measures of differential diagnostic value of diagnostic procedures. J Clin Epidemiol 1996;49:1435-9.
11. Kraemer HC. Evaluating Medical Tests: Objective and Quantitative Guidelines. Newbury Park, California: Sage; 1992.
12. Gregory RJ. Psychological testing: History, principles, and applications. 5th ed. Boston, MA: Pearson; 2007.
13. Saini KK, Sehgal RK, Sethi BL. Evaluation of general classes of reliability estimators often used in statistical analyses of quasi-experimental designs. AIP Conf Proc. 2008;1052:58-62.
14. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006;59:1033-9.
15. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, *et al*. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. Int J Nurs Stud 2011;48:661-71.
16. Mattu GS, Perry TL Jr, Wright JM. Comparison of the oscillometric blood pressure monitor (BPM-100(Beta)) with the auscultatory mercury sphygmomanometer. Blood Press Monit 2001;6:153-9.
17. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8:135-60.
18. Yoo JH, Kim JW, Kim KK, Kim JY, Koh BH, Lee EJ. Sasangin diagnosis questionnaire: Test of reliability. J Altern Complement Med 2007;13:111-22.
19. Abramson JH. Surveys Methods in Community Medicine. New York: Churchill Livingstone; 1990. p. 138.
20. Jang E, Baek Y, Park K, Lee S. Could the Sasang constitution itself be a risk factor of abdominal obesity? BMC Complement Altern Med 2013;13:72.
21. Kurande VH, Waagepetersen R, Toft E, Prasad R, Raturi L. Repeatability of pulse diagnosis and body constitution diagnosis in traditional indian ayurveda medicine. Glob Adv Health Med 2012;1:34-40.
22. Kurande VH, Waagepetersen R, Toft E, Prasad R Reliability of pulse diagnosis in traditional indian ayurveda medicine. In: 8th Annual Congress of the International Society for Complementary Medicine Research (ISCMR). Res Complement Med/Forsch Komplementmed 2013;20 Suppl 1:1-9.
23. Kim M, Cobbin D, Zaslawski C. Traditional Chinese medicine tongue inspection: An examination of the inter-and intrapractitioner reliability for specific tongue characteristics. J Altern Complement Med 2008;14:527-36.
24. O'Brien KA, Abbas E, Movsessian P, Hook M, Komesaroff PA, Birch S. Investigating the reliability of Japanese toyohari meridian therapy diagnosis. J Altern Complement Med 2009;15:1099-105.
25. Wickström G, Bendix T. The "Hawthorne effect" – What did the original Hawthorne studies actually show? Scand J Work Environ Health 2000;26:363-7.
26. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Meth 1996;1:30e46.
27. Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions. 3rd ed. Hoboken, NJ: Wiley; 2003.
28. Viera AJ, Garrett JM. Understanding interobserver agreement: The kappa statistic. Fam Med 2005;37:360-3.
29. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
30. Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213-20.
31. Altman DG. Practical Statistics for Medical Research. London: Chapman and Hall; 1996.
32. Brar BS, Chhibber R, Srinivasa VM, Dearing BA, McGowan R, Katz RV. Use of Ayurvedic diagnostic criteria in Ayurvedic clinical trials: A literature review focused on research methods. J Altern Complement Med 2012;18:20-8.
33. Sharma AK, Kumar R, Mishra A, Gupta R. Problems associated with clinical trials of ayurvedic medicines. Braz J Pharmacogn 2010;20:276-81.
34. Bhushan P, Kalpana J, Arvind C. Classification of human population based on HLA gene polymorphism and the concept of Prakriti in Ayurveda. J Altern Complement Med 2005;11:349-53.
35. Prasher B, Negi S, Aggarwal S, Mandal AK, Sethi TP, Deshmukh SR, *et al*. Whole genome expression and biochemical correlates of extreme constitutional types defined in Ayurveda. J Transl Med 2008;6:48.
36. Rastogi S. Development and validation of a Prototype Prakriti Analysis Tool (PPAT): Inferences from a pilot study. Ayu 2012;33:209-18.
37. Jang E, Kim JY, Lee H, Kim H, Baek Y, Lee S. A study on the reliability of sasang constitutional body trunk measurement. Evid Based Complement Alternat Med 2012;2012:604842.
38. Ryu H, Lee H, Kim H, Kim J. Reliability and validity of a cold-heat pattern questionnaire for traditional Chinese medicine. J Altern Complement Med 2010;16:663-7.
39. Dhruva A, Adler S, Weaver J, Acree M, Miaskowski C, Abrams D, *et al*. Mixed methods approaches in whole systems research: a study of ayurvedic diagnostics. BMC Complement Altern Med 2012;12(Suppl 1):378.
40. King E, Cobbin D, Walsh S, Ryan D. The reliable measurement of radial pulse characteristics. Acupunct Med 2002;20:150-9.
41. Kim J, Han GJ, Choi BH, Park JW, Park K, Yeo IK, *et al*. Development of differential criteria on tongue coating thickness in tongue diagnosis. Complement Ther Med 2012;20:316-22.
42. MacPherson H, Thorpe L, Thomas K, Campbell M. Acupuncture for low back pain: Traditional diagnosis and treatment of 148 patients in a clinical trial. Complement Ther Med 2004;12:38-44.
43. Zell B, Hirata J, Marcus A, Ettinger B, Pressman A, Ettinger KM. Diagnosis of symptomatic postmenopausal women by traditional Chinese medicine practitioners. Menopause 2000;7:129-34.
44. Sung JJ, Leung WK, Ching JY, Lao L, Zhang G, Wu JC, *et al*. Agreements among traditional Chinese medicine practitioners in the diagnosis and treatment of irritable bowel syndrome. Aliment Pharmacol Ther 2004;20:1205-10.
45. 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 medicine acupuncturists. J Altern Complement Med 2006;12:863-72.
46. Prlic HM, Lehman AJ, Cibere J, Sodhi V, Varma S, Sukumaran T, *et al*. Agreement among Ayurvedic practitioners in the identification and treatment of three cases of inflammatory arthritis. Clin Exp Rheumatol 2003;21:747-52.
47. Miller TP, Brennan TA, Milstein A. How can we make more progress in measuring physicians' performance to improve the value of care? Health Aff (Millwood) 2009;28:1429-37.

48. Patwardhan K, Gehlot S, Singh G, Rathore HC. Global challenges of graduate level Ayurvedic education: A survey. Int J Ayurveda Res 2010;1:49-54.
49. Furst DE, Venkatraman MM, McGann M, Manohar PR, Booth-LaForce C, Sarin R, *et al*. Double-blind, randomized, controlled, pilot study comparing classic ayurvedic medicine, methotrexate, and their combination in rheumatoid arthritis. J Clin Rheumatol 2011;17:185-92.
50. Tuijn S, Janssens F, Robben P, van den Bergh H. Reducing interrater variability and improving health care: A meta-analytical review. J Eval Clin Pract 2012;18:887-95.
51. Zhang GG, Lee WL, Lao L, Bausell B, Berman B, Handwerger B. The variability of TCM pattern diagnosis and herbal prescription on rheumatoid arthritis patients. Altern Ther Health Med 2004;10:58-63.
52. Zhang GG, Lee W, Bausell B, Lao L, Handwerger B, Berman B. 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;11:415-21.
53. Zhang GG, Singh B, Lee W, Handwerger B, Lao L, Berman B. 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;14:381-6.
54. Perloff D, Grim C, Flack J, Frolich ED, Hill M, McDonald M, *et al*. Human blood pressure determination by sphygmomanometry. Circulation 1993;88:2460-7.

How to cite this article: Kurande VH, Waagepetersen R, Toft E, Prasad R. Reliability studies of diagnostic methods in Indian traditional Ayurveda medicine: An overview. J Ayurveda Integr Med 2013;4:67-76.

Source of Support: "Erasmus Mundus Mobility for Life" Scholarship by European Commission at Aalborg University, Denmark for the first author only, **Conflict of Interest:** None declared.