Reliability of Diagnostic Methods in Indian Traditional Ayurvedic Medicine

Reliability of Diagnostic Methods in Indian Traditional Ayurvedic Medicine

PhD Thesis by

Vrinda Kurande

Clinical Science and Biomedicine,
Department of Health Science and Technology,
Aalborg University, Denmark

River Publishers
Aalborg

ISBN 978-87-93102-43-9 (e-book)

Published, sold and distributed by:
River Publishers
Niels Jernes Vej 10
9220 Aalborg Ø
Denmark

Tel.: +45369953197
www.riverpublishers.com

TABLE OF CONTENTS

ABBREVIATIONS

ATDS: Automated tongue diagnosis system

κ : Kappa

LK scale: Landis and Koch scale

SDQ: Sasangin Diagnosis Questionnaire

SM: Sasang medicine

TCM: Traditional Chinese Medicine

TMT: *Toyohari* meridian therapy

SUMMARY IN ENGLISH

In Ayurveda, body constitution, tongue, and pulse examination all have a long history of use; however, there is still a lack of quantitative data on the reliability of these diagnostic methods. In this context, reliability refers to maintaining consistency of information, a critical factor because consistent diagnoses lead to consistent treatment and are important for clinical practice, education, and research. Thus, the aim of the present PhD study has been to assess the reliability of Ayurvedic diagnostic methods. Cohen's weighted kappa statistic was used as a measure of reliability. Permutation tests were used to test the hypothesis of homogeneous diagnosis (i.e., the doctor's diagnosis does not depend on the subject).

The aim of study I was to assess the repeatability of pulse and body constitution assessment. A double-blind controlled study was conducted in Copenhagen. An Ayurvedic expert examined the pulses and body constitutions of 17 healthy participants twice in a random order without seeing them. A matrix of pulse and body constitution variables was developed. Moreover, it was discussed how the magnitude of the weighted kappa statistic may be interpreted using *p*-values calculated from random permutations of the data. The hypothesis of homogeneous classification was rejected on the 5% significance level (*p*-values of 0.02 and 0.001, respectively, for pulse and body constitution assessment). According to the LK scale, values of the weighted kappa for pulse examination ($\kappa = 0.42$) and body constitution assessment ($\kappa = 0.65$) correspond to "moderate" and "substantial" agreement, respectively. A reasonable level of consistency between the two pulse and body constitution assessments was observed.

The aim of study II was to assess the inter-rater and intra-rater reliability of pulse examination. In study II, 15 registered Ayurvedic doctors with 3-15 years of experience examined the pulses of 20 healthy subjects twice, making a total of 600 examinations. The

examinations were performed blind and in a random order. The weighted kappa statistics were negative for two doctors and ranged from 0.03 and 0.56 for the other doctors. Overall, there was very little evidence against the hypothesis of homogeneous diagnosis. The kappa values were generally larger in the group of experienced doctors (*p*-value 0.04) and course takers. Thus, experience and proper training seem to play a role in pulse examination

The aim of study III was to assess the inter-rater reliability of pulse, tongue and body constitution assessment. In this study, the same fifteen Ayurvedic doctors independently interviewed 20 healthy subjects and examined the body constitution, tongue, and pulse of each. Subjects completed self-assessment questionnaires and underwent software analysis for body constitution. Weighted kappa statistics for all 105 pairs of doctors were computed for the pulse, tongue, and body constitution data sets. According to the LK scale, the maximum pairwise kappas ranged from fair to moderate, slight to fair, and poor to slight for body constitution, tongue, and pulse assessment, respectively. For each data set and pair of doctors, the null hypothesis of random rating was rejected for just twelve pairs of doctors for body constitution, one pair of doctors for pulse examination and no pairs of doctors for tongue assessment. There was significant evidence against random software rating and the questionnaire that was used, and the diagnoses made by the majority of doctors. Body constitution assessment appears reliable when a questionnaire and software assessment is used, while other diagnostic methods have room for improvement.

The objective of study IV was to provide information about how the reliability studies can be designed and conducted for Ayurvedic diagnostic methods. In study IV, a review of literature was provided in order to illustrate relevant concepts of reliability studies of diagnostic methods and their implication in practice, education, and training. An introduction to reliability estimates, different study designs, and statistical analysis is given for future studies in Ayurveda.

In conclusion, this is the first study on the reliability of diagnostic methods in Ayurveda. The results showed that there is need for standardization and proper training to improve the reliability of diagnostic methods. The developed bio-statistical methodology might be beneficial for further reliability studies in Ayurveda.

DANSK RESUMÉ

Inden for Ayurveda har vurdering/bedømmelse af puls, tunge og kroppens konstitution været brugt igennem historien til diagnosticering. Der er mangel på kvantitative mål for pålideligheden af disse diagnostiske metoder. Pålidelighed betyder sammenhæng mellem oplysninger. Konsistente diagnoser fører til konsistente behandlinger, hvilket er vigtigt for klinisk praksis, uddannelse og forskning. Formålet med PhD studie er at vurdere pålideligheden af de ayurvediske diagnosticeringsmetoder. Cohens vægtede kappa statistik blev anvendt som pålidelighedsmål. Der blev anvendt permutationstests til at afprøve hypotesen for homogen diagnose (dvs. at lægens diagnose ikke afhænger af forsøgspersonen).

Formålet med Studie I var at vurdere reproducerbarheden af målingerne af puls og kropsforfatning. Studie I, som var et blindet, kontrolleret studie, blev udført i København. En Ayurveda-ekspert undersøgte to gange puls og kropskonstitution hos 17 raske deltagere i tilfældig rækkefølge uden at se dem. Der blev udviklet en skala for puls- og kropskonstitutionsvariabler. Hypotesen om homogen klassifikation blev afvist ved 5% af signifikansniveauet (*p*-værdier på henholdsvis 0,02 og 0,001 for puls og kropskonstitution). I henhold til Lands og Kochs skala svarer værdierne af vægtet kappa for undersøgelse af puls (κ = 0,42) og vurdering af kroppens konstitution (κ = 0,65) til henholdsvis ”moderat” og ”betydelig” overensstemmelse. Der observeredes et fornuftigt niveau af overensstemmelse mellem to vurderinger af puls- og kropsforfatning.

Formålet med Studie II var at vurdere inter-rater og intra-rater pålideligheden af pulsundersøgelsen. I Studie II undersøgte 15 registrerede Ayurveda-læger med 3-15 års erfaring, to gange pulsen på 20 raske forsøgspersoner; i alt 600 undersøgelser. Undersøgelserne blev udført blindet og i tilfældig rækkefølge. Den vægtede kappa-statistik var negativ for to læger og svingede fra 0,03 til 0,56 for de andre læger. Overordnet var der meget lav evidens for hypotesen om homogen diagnose. Kappa-værdierne var generelt højere i gruppen af erfarne læger (*p* 0 0,04). Endvidere var omfanget af vurderinger signifikant forskellige mellem lægerne (*p*-<0,0005).

Formålet med Studie III var at vurdere interrater-påligheden af målingerne af puls, tunge og kropsforfatning. I Studie III foretog de samme 15 Ayurveda-læger uafhængige interviews af 20 raske forsøgspersoner og undersøgte kropskonstitution, tunge og puls. Forsøgspersonerne udfyldte selvevalueringsskemaer og fik lavet en software-analyse af kroppens konstitution. Vægtet kappa-statistik for alle 105 par af læger blev udregnet for datasættene for puls, tunge og kropskonstitution. I henhold til Landis-Koch-skalaen svingede de maksimale parvise kappa fra ”rimelig” til ”moderat”, ”let” til ”rimelig” og fra ”ringe” til ”ubetydelig” for vurderingen af henholdsvis kropskonstitution, tunge og puls. For hvert datasæt og hvert par af læger blev nulhypotesen af den vilkårlige rating forkastet for kun tolv par af læger for kropskonstitution, et par af læger for puls-vurdering og ingen par af læger for tunge-vurdering. Således viser resultaterne et lavt niveau af pålidelighed for alle typer af vurderinger udført af lægerne. Der var signifikant evidens mod vilkårlig rating ved hjælp af den anvendte software og spørgeskemaerne og den diagnose, som hovedparten af lægerne foretrak. Vurdering af kropskonstitution synes pålidelig, når der anvendtes spørgeskema og software, mens andre diagnostiske metoder har plads til forbedring.

Formålet med Studie IV var at tilvejebringe informationer om, hvordan pålidelighedsstudier kan designes og udføres for ayurvediske diagnosticeringsmetoder. I Studie IV blev der tilvejebragt en litteraturgennemgang for at illustrere relevante koncepter af pålidelighedsstudier for diagnostiske metoder og deres betydning i praksis, undervisning og uddannelsesforløb. En introduktion til pålidelighedsestimater, forskellige studiedesigns og statistisk analyse er nødvendig for fremtidige studier i Ayurveda.

Afslutningsvis skal det nævnes, at dette er det første studie af pålideligheden af de diagnosticeringsmetoder, der anvendes i Ayurveda. Resultaterne viste, at der er et behov for standardisering og korrekt uddannelse for at forbedre pålideligheden af de diagnostiske metoder. Endvidere kunne den udviklede bio-statistiske metodologi være til nytte ved fremtidige studier af pålideligheden af Ayurveda.

PREFACE

This Ph.D. thesis is based on the following four research papers, which will be referred to in the text by their Roman numerals.

I. 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(5): 34-40.

II. Kurande VH, Waagepetersen R, Toft E, Prasad R. "Intra and Inter-rater Reliability of Pulse Examination in Traditional Indian Ayurvedic Medicine." *Integrative Medicine Research*.2013; 2(3): 89-98. .

III. Kurande VH, Bilgrau AE, Waagepetersen R, Toft E, Prasad R. "Inter-rater Reliability of Diagnostic Methods in Traditional Indian Ayurvedic Medicine." *Evidence-based Complementary and Alternative Medicine*. Volume 2013, http://dx.doi.org/10.1155/2013/658275,In Press.

IV. Kurande VH, Waagepetersen R, Toft E, Prasad R. "Reliability Studies of Diagnostic Methods in Indian Traditional Ayurveda Medicine – an Overview." *Journal of Ayurveda and Integrative Medicine*. 2013; 4(2): 67-76.

ACKNOWLEDGEMENTS

This Ph.D. thesis would not have been possible without the support of many people. First of all, I would like to express my sincere gratitude to the Director of the CTIF (Center for TeleInFrastructure), Professor Ramjee Prasad, who offered me the invaluable opportunity to do research at Aalborg University, Denmark. I would also like to express my deep thanks to my supervisor, Professor Egon Toft, Dean of the Medical Faculty, for his enthusiastic supervision, inspiration, and encouragement. I would like to thank Professor Rasmus Waagepetersen from the Mathematical Department for the statistical help.

I would also like to thank all of my colleagues, especially Dipashree Bhalerao, Kirti Mehendale, Anders Ellern Bilgrau, Nitika Sachdeva, Kirti Pasari, Shinichiro Shiozawa, and Swati Prasad. The doctors and the volunteers who participated in my experiments also deserve acknowledgement. Finally, I wish to convey my heartfelt thanks to my wonderful family: my father Hari Kulkarni, my mother Vinaya, my beloved husband Hitendra, and my son Aditya, for their long-lasting patience and unconditional support.

The present Ph.D. study has been financially supported by the “Erasmus Mundus Mobility for Life” Scholarship, which is provided by the European Commission at Aalborg University, CTIF, Denmark. Additionally, parts of the research presented at the seventh and eighth International Congresses for Complementary Medicine Researchers in Portland, Oregon, USA, 2012 and London, UK, 2013 were supported by Department of Health Science and Technology, Aalborg University, Denmark.

1. BACKGROUND

1.1. Introduction

Traditional Indian Ayurvedic medicine has been practiced in the Indian peninsula for more than 5000 years. This medical system was established from 2500 to 600 BC [1]. The word “Ayurveda” is derived from the Sanskrit word, “*Ayus*” and “*Veda*”. “*Ayus*” (life) stands for the combination of the body, the mind, the sense organs, and the soul. “*Veda*” means science or knowledge. Its main objective is to achieve optimal health and wellbeing through a comprehensive approach that addresses mind, body, behavior, and environment. This holistic system of medicine emphasizes prevention and health promotion, and provides treatment for diseases [2, 3, 4, 5]. Ayurveda is widely used in India as a system of primary health care. Presently, there are 2420 hospitals, 429,246 registered practitioners, more than 320 educational institutions, and 7699 drug-manufacturing units to promote Ayurveda in the health care delivery system in India [6]. In the West, it is popular as an alternative or complementary medicine, and it has been used in the form of herbal medicine, yoga, dietary supplements, and massage. It is widely used in North America, Canada, Australia, and the UK and has spread beyond the ethnic populations from which the traditional medicine practices originated [7]. The ancient concept of evidence of Ayurveda is based on epistemology and logic (*pramana vijnan*), meaning direct observation (*pratyaksha praman*), inferential evidence (*anuman*), scriptural evidence (*Aptopadesha*), and rationally planned experimental evidence (*yukti Praman*). In contemporary Ayurveda, established tradition and scriptural evidence have an important role [8]. Now, healthcare relies on the best evidence available in making a decision about the patient’s wellbeing. The Western approach relies on the scientific evidence of any phenomenon [9], and as in the development of other branches of knowledge, contemporary Ayurveda seeks supportive scientific evidence in today’s post-modern era [10]. Thus, using a more scientific approach will not only help in updating current knowledge, but also help Ayurveda to gain global exposure. Currently, few efforts are being made to search for scientific evidence for Ayurveda, but this trend is still in the preliminary stage and is facing many challenges [11 -13].

1.2. Ayurvedic Philosophy

According to the fundamental principle of Ayurveda, everything that exists in the macrocosm also appears in the internal cosmos of the human body (microcosm). Consequently, the universe is governed by the wind, the sun, and the moon; similarly, human body and mind are also governed by three *doshas* (body humors) *vata*, *pitta*, and *kapha*, respectively (Appendix). The entire macrocosm and microcosm is made up of five great elements: earth, water, fire, air, and ether. Thus, *vata* is the combination of air and space, *pitta* is made up of fire and water, and *kapha* is the combination of earth and water [3, 14].

The human body is combination of:

- Three *doshas* and sub-*doshas.*
- Seven *dhatus* – fundamental principles that support the various bodily tissues; these are chyle (*rasa*), blood (*rakta*), flesh (*mamsa*), fat (*meda*), bone (*asthi*), marrow (*majja*) and reproductive tissues (*shukra*).
- Three *malas* - the body’s excretory products are feces (*purisha*), urine (*mutra*), and sweat (*sweda*).
- 13 srotas (channels of the body).

The *dhatus* of the body are the structural entities whereas *doshas* are physiological entities, derived from different combination and permutations of basic elements. The growth and decay of this matrix and its constituents revolve around food which gets process into humors, tissues and wastes. Ingestion, digestion, absorption, assimilation and metabolism of food have interplay in health and disease which are significantly affected by psychological mechanisms as well as by bio- fire (*agni*). The food/drug is composed of five elements, which replenish or nourish the like element of the body after the action of bio fire.

Ayurveda classify disease cause in

- Diet (*Aaharaja*)
- Activity related (*Viharaja*)
- Psychological (*Manasa*)

Disease begins when person is living out of harmony with his own environment. Inappropriate impression from environment disturbs the internal environment and cause disharmony. In Ayurveda, state of perfect health is balanced status of these three *doshas*. Perfect balance of theses *doshas* means when all of the *doshas* are functioning correctly in correct proportion to each other and in their right locations. Imbalance of these *doshas* means some of this aspect is not satisfied and hence manifest as a disease. This process of imbalance is explained in 6 stages; accumulation, aggravation or vitiation, spreading or migration, condensation or localization, manifestation and an acute crisis [15].

Stage I: Accumulation—a disproportion of a particular *dosha* or *subdosha* increases or accumulates.

Stage II: Aggravation or Vitiation—the *dosha* or *subdosha* becomes overstimulated and may go into a wrong channel. The *dosha* or *subdosha* begins to act in an inappropriate way.

Stage III: Spreading or Migration—the *dosha* or *subdosha* moves beyond the location of its normal function.

Stage IV: Condensation or Localization—the *dosha* or *subdosha* localizes in an inappropriate region.

Stave V Manifestation—Pathologic symptoms manifest in the new location, e.g. a boil, arthritis or angina.

Stave VI: Bursting (possibly)—an acute crisis occurs as with an aneurysm, an embolism or an infarction.

Ayurvedic differential diagnosis must understand which *dosha* has been involved in the disease formation, process and different stages of the disease.

1.3. Diagnosis in Ayurveda

Thus, diagnostic decision making in Ayurveda is a complex process. In Western medicine, the continuous evolution of biomedical methods has influenced conventional decision making. In Ayurveda, however, decision making depends only upon diagnosing the disease through subjective and objective clinical examination. Ayurvedic methods of diagnosis and diagnostic classifications are performed principally to find the root cause of the disease and then select the appropriate treatment. The main task of the Ayurvedic physician is to diagnose the imbalance of one or several *doshas*, and treatment includes reestablishing the proper balance [14]. In Ayurveda, the diagnosis extends beyond the limit of an objective description of what is wrong with the patient. It envelops all anatomical structures; for instance, bodily constituents (*dhatu-s*), excretory products (*mala-s*), digestive power (*agni*), and body channels (*srota-s*), all of which can be involved in disease manifestation. It also considers pathogenic factors, season of the year, and the entire course of action (diet, drug and regimen compatible to the constitution) that may have led to development of the disease. Recently, In Western medicine, various diagnostic tests have shown to be of aid in refining diagnoses, detecting diseases, and providing prognostic information. Conversely, Ayurvedic diagnosis depends on the physician's knowledge of authoritative medical texts, his/her observation of physical signs, the patient's narrative, and valid inferences.

1.4. Diagnostic Methods

Diagnosis is essentially done by following three methods [2];

- Observation (*Darshana*)
- Touch/Palpation (*sparshana*)
- Questioning /interrogation (*prashna*)

The authentic texts explain the use of five senses while examining a patient and asking the relevant questions. The physician uses a bimodal approach (the disease diagnosis and a diagnosis of the patient's overall health) of clinical examination (Figure 1). The classical clinical examination involved eight fold examination (*ashta sthana pariksha*); pulse, urine, faeces, tongue, voice, skin or touch (*sparsha*), eyes (*druk*), the overall appearance of a patient. This eight fold examination is necessary to identify deranged *doshas*. In addition to these, tenfold examination is done to determine the best combination of therapy and dosage to minimize the adverse effect of the drug. These tenfold examinations are tissues vitality, habitation (*desha*), body strength, body constitution, season and disease condition, digestive capacity, age, mental nature, adaptability, and diet.

Disease examination involves etiology, warning symptoms, symptoms, therapeutic measure for diagnostic purpose, and pathogenesis (disease stages). Importance of diagnosis in the practice of Ayurveda gives it a logical priority; a correct diagnosis will pave way for prognosis and treatment.

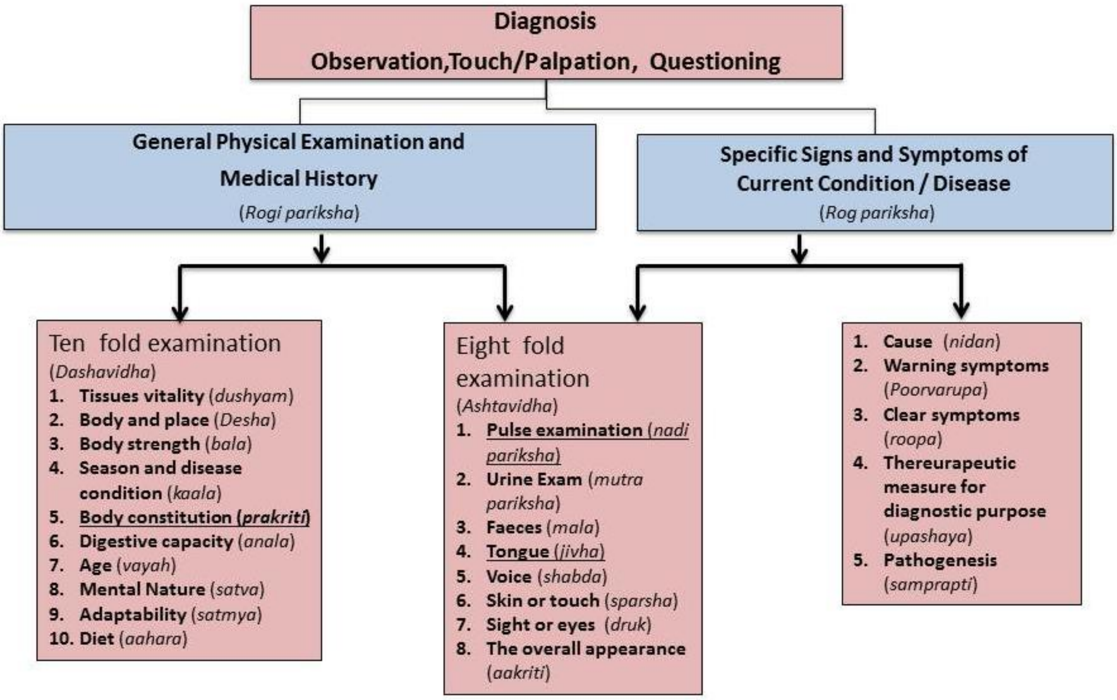

Diagnosis
Observation,Touch/Palpation, Questioning
General Physical Examination and Medical History
(Rogi pariksha)
Specific Signs and Symptoms of Current Condition / Disease
(Rog pariksha)
Ten fold examination
(Dashavidha)
1. Tissues vitality (dushyam)
2. Body and place (Desha)
3. Body strength (bala)
4. Season and disease condition (kaala)
5. Body constitution (prakriti)
6. Digestive capacity (anala)
7. Age (vayah)
8. Mental Nature (satva)
9. Adaptability (satmya)
10. Diet (aahara)
Eight fold examination
(Ashtavidha)
1. Pulse examination (nadi pariksha)
2. Urine Exam (mutra pariksha)
3. Faeces (mala)
4. Tongue (jivha)
5. Voice (shabda)
6. Skin or touch (sparsha)
7. Sight or eyes (druk)
8. The overall appearance (aakriti)
1. Cause (nidan)
2. Warning symptoms (Poorvarupa)
3. Clear symptoms (roopa)
4. Thereuapeutic measure for diagnostic purpose (upashaya)
5. Pathogenesis (samprapti)

Figure 1: Diagnosis in Ayurveda: an overview [14]

1.5. Treatment Principles

In Ayurveda, meaning of treatment (*chikitsa - Kit Rogapanayane*) means disease diagnosis and management both from preventive, and curative aspects till their recurrence is sealed off. It includes all such measures that will establish equilibrium of body competent and re-establish health. Thus, it includes not only medication (herbs, minerals, metals and compound formulae) but also other non-material methods e.g. yoga, meditation, deities, counseling etc. [2, 4, 14]. Ayurvedic management is special in the sense that it is holistic in every sense and ensures non-recurrence of the disease and disease specific rejuvenation. Thus, the aim of the Ayurvedic treatment is:

- Re-establish the deranged balance of *doshas*
- Cleanse the contaminated *dhatus* and other constituents.
- Improve and maintain *agni* (digestive fire) activity at optimum.
- Maintain smooth function of mind, sense organs and thus establish an optimal state of function of all the body constituents and induce a sense of wellbeing.

Health is not mere absence of disease but a positive state, which needs to be maintained.

1.6. Treatment Methods of Ayurveda

Different type treatment can be administered based on patient's diagnosis [3, 4, 14].

- *Shodhana* (cleansing therapy):-five procedures that physically eliminate the contaminant, vitiated *doshas: emesis*, purgation, enemata, nasal instillation, and blood-letting). *Snehan* (body massage) and *swedan* (fomentation) procedures are prerequisite for cleansing therapy.
- *Shaman* (Palliative measures): Instead of elimination of vitiated *doshas* these methods render vitiated *doshas* by improving individual's digestive capacity (*agni*). Thus, *doshas* do not remain contaminated any more. These methods are:
  - *Pachana* - Intensify the digestive activity

- *Deepana*- intensify the gastric fire
- *Kshut nirodha* – withhold hunger, fasting, rest to the overwork digestive system
- *Tritnigrah* – withhold thirst
- *Vyayam*- physical activity
- *Atapa seva* – Exposure to sun's rays
- *Maarut sevaa* – exposure to dry clean wind or clime, or breathing technique
- *Nidan parivarjana* -avoidance of disease causing and aggravating factors
- *Pathya* – Diet and activity plan

- *Rasaayana* and *vajikarana*- it is rejuvenation branch of Ayurveda. Specific to the disease, age and seasons. *Vajikarana* deals mainly with the reproductive tissue metabolism as well as regenerative activity / capacity of the body.
- Other types of treatment
  - *Satvavajay* - mental nurturing and spiritual healing. This involves psychological measures like counseling and use of antagonistic emotions.
  - *Yuktivyapashray* – control the condition logically and rationally
  - *Antah parimaarjana*- group of modalities that when used internally manage to establish equilibrium of the vitiated *doshas*
  - *Bahih Parimaarjana* – external cleansing e.g. massage, anointing, sprinkling of medicated oil or liquids on the affected path.
  - *Shastra pranidhaana*- surgical intervention.

The treatment is often accompanied or followed by oral administration of Ayurvedic herbal formulas two to three times per day. These herbs may be administered in tablet, powder, juice or decoction forms. Metal and mineral preparations called *bhasma* are used extensively in Ayurvedic medicine.

Despite Ayurveda’s comprehensive foundation of diagnostic methods, there is a lack of scientific evidence and quantitative studies on the reliability of these diagnostic methods as evaluated by modern medicine. Validity and reliability are essential requirements of all outcome measures. [16].

1.7. Reliability Concept

In research, the reliability of diagnostic methods refers to the overall consistency with which they render a specific diagnosis [17]. A diagnostic method is said to be reliable if it produces similar results under identical conditions. This is a very important property of any diagnostic method in any clinical trial or practice because it is essential to establish that any changes observed in a patient/individual are due to the physician's intervention rather than problems in the diagnostic methods.

Several general types of reliability estimates are explained in (Table 1): (Study IV)

Table 1: Types of reliability

Types of reliability	**Methods/classification/ instrument**	**Raters**	**Subjects**	**Time / settings**
Intra–rater reliability	Same	Same	Same	Different
Inter–rater reliability	Same	Different	Same	Same
Inter–method reliability	Different	Same	Same	Different or same
Internal consistency –reliability	Different items on the same test			

Reliability is concerned with the reproducibility or repeatability of a certain outcome. Intra-rater reliability more directly evaluates whether a method yields the same result upon repeated application by the same raters on the same subjects. It is also known as test-retest reliability or repeatability. Second assessments need to be taken after a length of time sufficient to ensure that raters are unlikely to recall their previous diagnosis, but not so long that actual changes in the original diagnosis have occurred. The usual range of time elapsed between assessments tends to be between two and fourteen days [18]. However, in some quickly changeable signs and symptoms like pulse examination, reassessment needs to be performed in the same setting within a short period of time to prevent changes in the characteristics observed by the raters. In this situation, blinding can be imposed to avoid any potential carry-over of the previous assessment.

In Ayurveda, physical examination findings such as pulse examination and body constitution often rely on some degree of subjective interpretation by doctors. If the doctors who interpret the diagnosis cannot agree on the interpretation, the results will be of little use. Inter-rater reliability is the degree to which two or more raters are able to differentiate among subjects under identical assessment conditions. Reliability results are important issues in classification, scale and instrument development, quality assurance, and in the conduct of clinical studies [19].

2. LITERATURE REVIEW

Several reliability studies are conducted in western medicine [20]. The investigation of the reliability of traditional Chinese medicine (TCM), Toyohari meridian therapy (TMT), and Sasang Constitutional Medicine (SCM) diagnoses is in the formative stage [21, 22]. However, reliability studies in Ayurveda are in the preliminary stage. In the literature review, we focused on the reliability studies carried out in TCM, TMT, and SCM. These traditional medicines have some similarities with Ayurveda. One 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. Thus, in study IV, examples are provided to illustrate relevant concepts of reliability studies of diagnostic methods from different traditional medicine.

Database search and information sources

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. [20-22] Articles were limited to those in the English language. The scope of this review was limited to inter and intra rater reliability for specific diagnostic methods: pulse examination, body constitution diagnosis and tongue diagnosis in all types of traditional medicine. Reliability studies in different traditional medicines from Asia are described in (Table 2), and reliability studies in Ayurveda are described in (Table 3).

Data items

The following information was extracted from each study: (1) Authors; (2) Diagnostic method; (3) Type of reliability; (4) Subject studied; (5) Observers; (6) Study design; (7) Statistical test; (8) Results.

Table 2: Reliability studies in different traditional medicines from Asia

Intra-rater reliability studies done in different traditional medicines from Asia						
Authors	**Diagnostic method**	**Subject studied**	**Observers**	**Study design**	**Statistical test**	**Results**
Jang, 2013 [23].	SC Sasang constitution	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
Yoo *et al*., 2007 [24].	Sasang constitution Questionnaire (SDQ)	88 questionnaires	Self-report structured questionnaire	SDQ was administer twice with an interval of 2 weeks	Pearson’s correlation coefficients	0.44-0.74
Kurosu 1969 [25] .		40 subjects, half healthy and half with a medical condition	17 meridian therapy practitioners	Blinded examination twice in a random order	No formal statistical analysis	Healthy subjects 43–65% (mean 53%) Patients: 50–75% (mean 60%). Average percentage = 55.8 %
Kim *et al*., 2008 [26].	TCM tongue inspection	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 [27].	TCT,	50 tongue photographs Digital tongue imaging system (DTIS)	24 reliable assessors were selected from 60 oriental medical doctors	Raters TCT judgments and Digital tongue imaging system (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)
Inter-rater reliability studies in different traditional medicines from Asia						
King *et al*., 2002 [28].	TCM, pulse diagnosis	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	Average 80%

O'Brien *et al*., 2009[29].	Pulse diagnosis, abdominal diagnosis and *sho* diagnosis in TMT	Sixty-two 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 [30].	TCM diagnosis on patients with RA, the tongue and pulse and a herbal prescription	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 [31].		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 [32].		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)

Table 3: Inter-rater reliability studies in Ayurveda

Authors	**Ayurvedic diagnostic method**	**Subject studied**	**Ayurvedic practitioners**	**Study design**	**Statistical test**	**Result**
Rastogi, 2012 [33]	Prototype prakriti analysis tool (PPAT)	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 [34]	Ayurvedic disease origin, diagnosis and treatment approach for inflammatory arthritis	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 [35].	*Prakriti* and *vikruti*	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%

Discussion of literature analysis

- The majority studies have investigated reliability of pulse examination, with results ranging from low to a good level of agreement. Studies of reliability of tongue diagnosis in a TCM reported 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 the reliability. Including practitioner variability due to differences in clinical education and experience [21].
- Many studies on inter-rater reliability included 2- 30 raters. Inter-rater reliability decreases as the number of raters increases. It is recommended that to extensively test inter rater reliability the number of raters that evaluate each subject must be greater than two [21, 22].

- For intra- rater reliability study, proper time interval is necessary to avoid the carryover effect of the first diagnosis in case of body constitution diagnosis as shown in [23,24]. However, it is not possible for pulse and tongue diagnosis in Ayurveda. As pulse characteristics will change within an hour. 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. Furthermore, blinding and randomization is necessary to avoid carry over effect of the previous diagnosis [25]. For tongue diagnosis, realistic slides may be one option. But the quality of these slides/photographs should be good [26].
- It is possible to develop evidence based guideline by reliable raters and using new technology. Accordingly, it is proposed that thick coating of the tongue is that occupying approximately more than two-third of the tongue surface area [27].
- Good levels of inter rater reliability are possible when the system of pulse examination is operationally defined [28].
- Improvement in the level of reliability was observed after training sessions for the practitioners [29].
- To ensure independent data the raters in the study must not have prior knowledge of the subjects' signs and symptoms [21].
- The literature review identified the lack of proper statistical methods in most of the studies. Hence, we have developed statistical method for data analysis.
- Reliability studies in Ayurveda are limited only to inter-rater reliability with lack of proper statistical methodologies. Consequently, there is need to comprehensively assess the reliability studies in Ayurveda.

3. OBJECTIVES

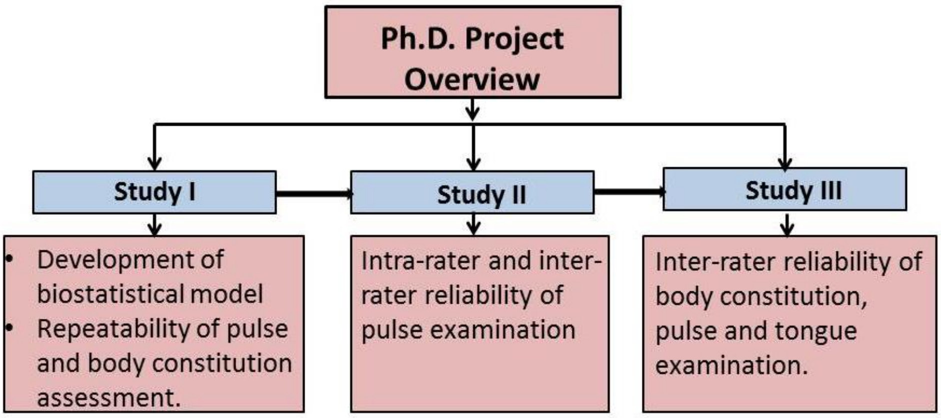
Adoption of unambiguous, reliable, and comprehensively applicable methods to generate evidence is a prerequisite for evidence-based decision making in Ayurveda. Thus, the objective of this Ph.D. study is to investigate the reliability of the diagnostic methods used in Ayurveda (Figure 2).

Study I assessed the intra-rater reliability (repeatability) of pulse and body constitution assessment to provide additional interpretation of Cohen's weighted kappa statistic for analysis of categorical diagnosis variables.

Study II tested the intra and inter-rater reliability of pulse examination.

Study III assessed the inter-rater reliability of body constitution, tongue, and tongue examination.

Study IV provided thorough literature review on the types of reliabilities, how to assess them, and similar studies conducted in Ayurveda and other traditional medicines from Asia.



Ph.D. Project Overview
Study I
Study II
Study III
• Development of biostatistical model
• Repeatability of pulse and body constitution assessment.
Intra-rater and inter-rater reliability of pulse examination
Inter-rater reliability of body constitution, pulse and tongue examination.

Figure 2: Ph.D. Project overview

4. STUDY METHODOLOGY

4.1. Diagnosis Outcome Variables

The Ayurvedic concepts of physiology, pathology, diagnosis, medicine, and therapeutics are based on the doctrine of the three *doshas*; the three bodily humors/energies/principles are known as *Vata, Pitta*, and *Kapha*—the subtle entities of the psycho-physiological body (Appendix 1). The three *doshas* are related to the modern scientific framework of systems theory, phase transitions, and irreversible thermodynamics. *Vata dosha* and its subtypes can be identified as regulating input/output processes and motion, *pitta* and its *subdoshas* can be identified as regulating throughput, turnover, and, hence, energy, and *kapha* and its *subdoshas* can be identified as regulating storage, structure, and lubrication [36].

Apart from disease diagnosis and prognosis, Ayurvedic differential diagnosis must understand which *dosha* has been involved in the disease formation and process. Every *dosha* is proposed to have inherent attributes which are expressed in the physical, psychological, and physiological characteristics in an individual. *Charaka Samhita* explicitly explains how to identify *dosha* properties through signs and symptoms leading to the manifestation of body constitution and diseases [14]. Based on the *dosha* differential diagnosis, the most effective therapy for both the normal and diseased condition is prescribed. Thus, the first step in the diagnosis involves body constitution assessment. Understanding the proper natural combination of *doshas* typically guides the doctor in understanding the patient's shift from a state of balance to a state of disease.

4.2. Body Constitution Assessment Method

Prakriti refers to the consequences of the relative proportions of three *doshas*. In clinical practice, according to the characteristics of three *doshas*, the dominance of one or more *dosha* is assessed to classify individual subtypes by observation, touch, and questions.

Seven types of body constitutions exist, namely *vata, pitta, kapha, vatapitta, vatakaph, pittakapha*, and *tridoshaja* (the combination of the three *doshas*) [37]. Body constitution assessment is reached through an active–passive interaction between person and doctor, after which the doctor chooses the most appropriate classification. Recently, symptom-based checklists (questionnaires) have been developed to aid the interview process. However, there is currently no standardized questionnaire. Thus, in the present study, we prefer to assess the doctor-subject interaction instead of limiting the doctor's diagnosis by providing a questionnaire. Furthermore, each subject was asked to complete a self-report questionnaire and utilize the Ayusoft *prakriti* analysis software developed by the Centre for Development of Advanced Computing (C-DAC), Pune University, Pune, India.

4.3. Tongue Diagnosis

Tongue diagnosis is one part of an eightfold diagnostic method. In Ayurveda, the tongue is the mirror of the viscera. The presence of coating on the tongue is an indication of derangement of the digestive tract and the presence of *aama* (toxins) in the digestive organs (Appendix) [39]. In the present study, doctors were asked to diagnose presence (*saama jihva*), moderate presence (*alpa saama jihva*), or absence (*niram jihva*) of tongue coating.

4.4. Pulse Examination Method

In this study, the doctors examined the most commonly observed *nadi*, *jivanadi* (radial artery). The doctors placed their index finger below the radial styloid on the radial artery of the subject. The middle and ring fingers were placed next to the index finger. The pulse was taken from the left hand for the female participants and from the right hand for the male participants. An expert doctor is expected to be able to predict physiological condition, mental state, and general pathological state by pulse reading. However, the most important aspect of the pulse examination is to determine the qualities of *doshas* within the pulse. The patterns of pulse also depend on the level of *tridosha*. Natural qualities of *dosha,* like a snake’s curved scrawling under the index finger for *vata dosha*, a sensation like a frog jumping under the middle finger for *pitta dosha,* and a swan's smooth, slow movement felt under the ring finger for *kapha dosha,* are diagnosed during the pulse examination [40, 41, and 42]. In the present study, all practitioners were assumed to be using the same pulse qualities.

4.5. Study Subjects

The first study was conducted on 17 healthy subjects (males: n=2, females: n=15, age 18 to 60 years). Studies II and III were conducted on 20 healthy subjects (males: n=10, females: n=10, average 19±1 years), second year students from Sri Sri College of Ayurvedic Science & Research Hospital. An age of 18 years or older was an inclusion criteria. All subjects were in good health and no one was on medication. Exclusion criteria included: hunger or thirst, anointment with oil, having been recently awoken or bathed, diabetes, hypertension, and the presence of medications for anti-hypertension or hormone replacement therapy purposes. Written consent was obtained from all participants.

4.6. Ayurvedic Doctor

The first study was conducted by a registered practitioner and expert in pulse examination with more than ten years of practice. Studies II and III were conducted by 15 registered practitioners with different levels of clinical experience (Figure 3), all of whom have been practicing at Sri Sri College of Ayurvedic Science & Research Hospital. Ten doctors were MD in Ayurveda, two were MS in Ayurveda, and three were bachelors in Ayurveda. Among the 15 participating doctors, doctors 3, 5, 7, 10, and 11 have much more experience in this field than the others. Additionally, doctors 13, 14, and 15 completed a one month course in pulse examination at Sri Sri Ayurveda Trust. All of them are knowledgeable in pulse qualities and regularly use this method in combination with other diagnostic methods.

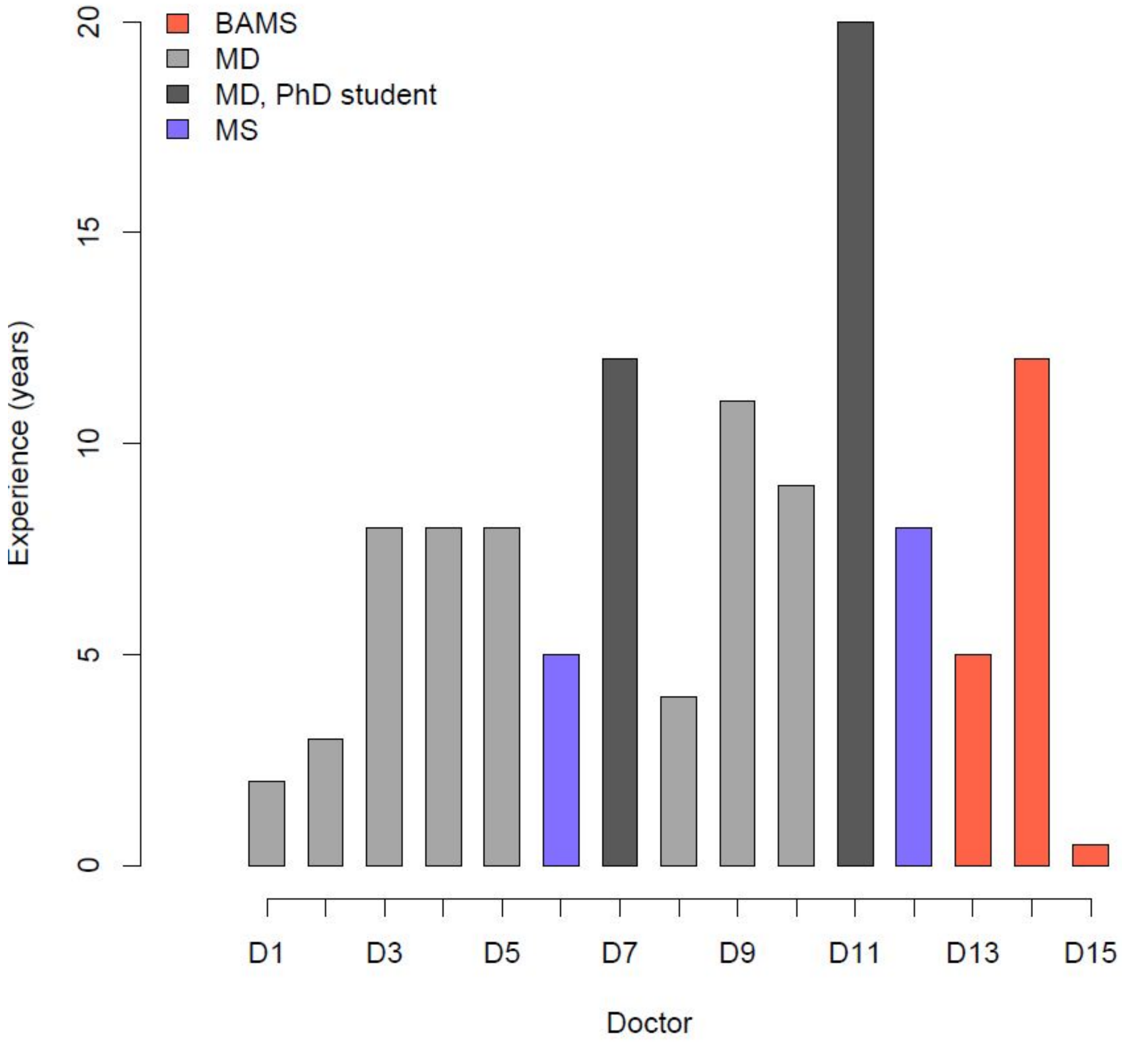


Figure 3: Ayurvedic Doctors' clinical experience and educational level

4.7. Study Procedure

Before the pulse examination, all participating subjects fasted for two hours for the first study. The study was conducted from 1:00 p.m. to 3:00 p.m. in the afternoon. The doctor examined each participant twice.

The second and third studies were conducted on the same day. The second was conducted in the morning and lasted from 10:00 a.m. to 11:30 a.m. Before pulse examination, all participants fasted for two hours. All doctors blindly and independently diagnosed only the pulse of each subject, twice. Then, the third study was conducted by each doctor interviewing (non-blinded) the subjects independently in order to assess body constitution, tongue, and pulse. The doctors were allowed to use all diagnostic methods, questioning, palpation, and observation, to arrive at the final diagnosis. Also, doctors were allowed to take breaks whenever necessary to avoid fatigue. After the examinations were completed, all subjects were given a week to independently complete the self-rating body constitution questionnaires and the Ayusoft *Prakriti* software analysis. To avoid a change in pulse pattern as a result of time of day differences or changes in circumstance, both exams for this portion of the study were conducted on the same day and in a short time period.

4.8. Randomization and Blinding

Study I and study II were blinded studies. Since the objective of study I was to investigate the repeatability of pulse examination, randomization and blinding was used to avoid a possible carry-over effect of the first diagnosis (Table 4). Blinding and randomization were implemented as follows: Subjects placed their arms (only palm and wrist) through a hole in a curtain separating the doctor from the subject. The doctor did not communicate with any of the participants. The participants entered in a random order, unknown to the doctor, and the doctor was unaware of the number of participants. Although the body constitution assessment was done based on deep pulse and hand observation, the doctor did not keep any record. Also, participants were asked to remove rings, wristwatches, and other hand ornamentation before showing their hands in order to make sure that the doctor could not easily recognize any of the participants on the second occasion.

The same method was applied for study II. In the second study, subjects also wore gloves.

Table 4: Study methods: an overview

	Type of diagnosis	Type of reliability	Design	Number of healthy subjects	Number of Ayurvedic doctors
Study I	Pulse and Body constitution	Intra-rater	Blinded controlled study	17	One (very experienced)
Study II	Pulse examination	Intra-rater and inter-rater	Blinded controlled study	20	15
Study III	Body constitution, tongue, and pulse examination	Inter-rater	Independent Controlled	20	15
	Body constitution	Method comparison	Independent Controlled	20	20 questionnaires and 20 software analyses

4.9. Ethical Consideration

Based on the description of study I, the Research Ethics Committee of North Jutland stated that the project could not covered by the science ethical treatment of medical research § 2 act. Thus, the project was not noticeable to the committee system and could be conducted from an ethical stand point. The same was applicable to study II and study III. In the first study, a lecture on pulse examination and Ayurveda was given to the subjects in the Art of Living Center on the previous day. All subjects were given information regarding the purpose of the study, its potential risks and benefits, and an explanation of data collection procedures and the time required for the experiment. Participation was voluntary and written consent was obtained from those agreeing to participate.

5. STATISTICAL METHODOLOGY

5.1. Comparison of Pulse and Body Constitution Assessment

The result of a pulse or body constitution assessment is a nominal variable and can be categorized into 10 classes, all corresponding to various mixtures of *vata*, *pitta* and *kapha* (Table 5, Figure 4). Regarding combinations of types, *vatapitta* means that *vata* is dominant but *pitta* is also present, while in *vatapittakapha*, all types are equally represented. When comparing pulse diagnoses, it is clear that, for example, class 1 (*vata*) is closer to classes 2 and 3 (*vatapitta* and *vatakapha*) than it is to classes 5 (*pittavata*) or 9 (*kaphavata*). For easy comparison of diagnoses, we have developed a distance measure which takes into account the proportions of the basic *dosha* types when comparing diagnoses.

Table 5: Diagnosis classes and weights for each variable

Diagnosis Classes	**Types of Pulse**	**Body Constitution (*Prakriti*)**	**Weights for each variable**
C_1	*Vata*	*Vataja*	(1,0,0)
C_2	*Vatapitta*	*Vatapittaja*	(2/3, 1/3,0)
C_3	*Vatakapha*	*Vatakaphaja*	(2/3, 0,1/3)
C_4	*Pitta*	*Pittaja*	(0,1,0)
C_5	*Pittavata*	*Pittavataja*	(1/3, 2/3,0)
C_6	*Pittakapha*	*Pittakaphaja*	(0, 2/3, 1/3)
C_7	*Kapha*	*Kaphaja*	(0,0,1)
C_8	*Kaphavata*	*Kaphavataja*	(1/3, 0, 2/3)
C_9	*Kaphapitta*	*Kaphapittaja*	(0, 1/3, 2/3)
C_{10}	*Vatapitta kapha*	*Vatapitta Kaphaja(tridoshaja)*	(1/3, 1/3,1/3)

5.2. Distance Measure on Diagnosis

In order to formalize mathematically that some pulse examination or body constitution assessments are closer than others, we assigned numerical weights which quantify the proportions of the three basic types, *vata*, *pitta,* and *kapha,* for the 10 classes (Table 3). In study I and III the distance between two classes c_1 and c_2 with weight vectors $w_1 = (w_{11}, w_{12}, w_{13})$ and $w_2 = (w_{21}, w_{22}, w_{23})$ is defined by the formula:

$$\mathrm{D}(c_1,c_2)=1-\frac{w_1}{\|w_1\|}\cdot\frac{w_2}{\|w_2\|}$$

This distance measure makes intuitive sense as explained below. It further corresponds to a sum of squared deviations similar to what is used in conventional analysis of variance. The minimal distance “0” occurs when the doctor diagnoses the same *dosha* the first and second time. The maximal distance “1” is obtained when two classes have none of the basic types (*vata*, *pitta*, or *kapha)* in common (Table 6). For instance, if the doctor diagnoses *vata* both the first and second time D(*vata*, *vata*) = 0. If the doctor diagnoses *vata* the first time and *vata-pitta* the second time, this means he is able to diagnose at least the dominant *dosha* and the two diagnoses overlap considerably. In this case, D (*vata*, *vatapitta*) = 0.11 (Figure 4).

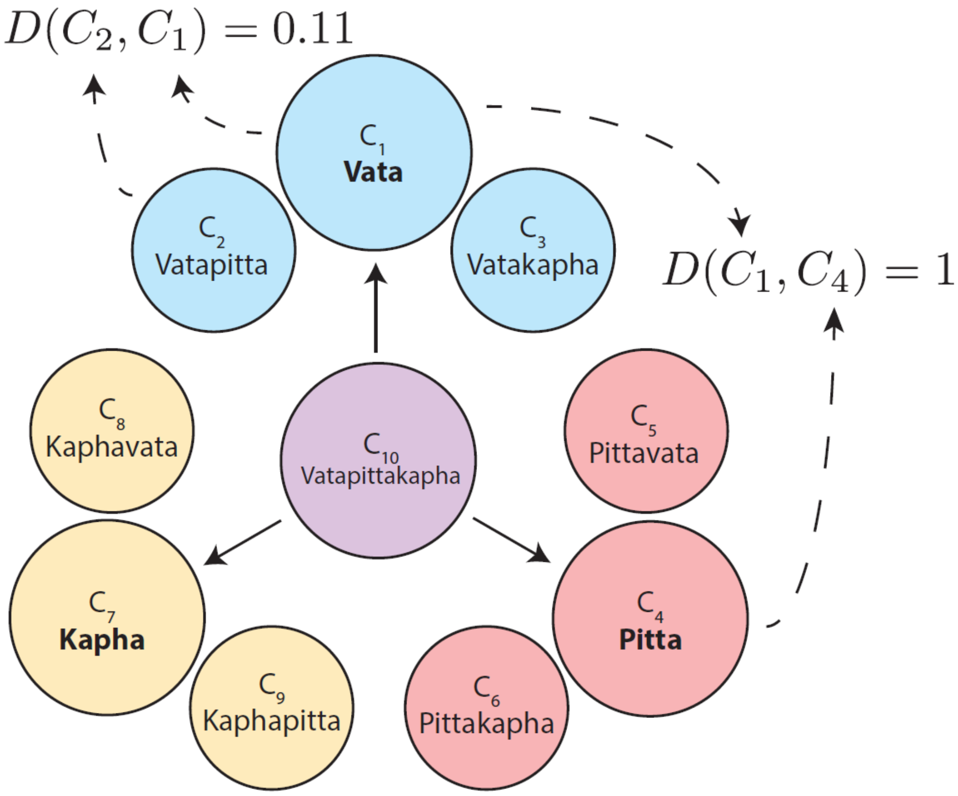


Figure 4: Distance between two classes

Table 6: Distance matrix between two classes

Classes	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
C_1	0	0.11	0.11	1	0.55	1	1	0.55	1	0.42
C_2	0.11	0	0.2	0.55	0.2	0.6	1	0.6	0.8	0.23
C_3	0.11	0.2	0	1	0.6	0.8	0.55	0.2	0.6	0.23
C_4	1	0.55	1	0	0.11	0.11	1	1	0.55	0.42
C_5	0.55	0.2	0.6	0.11	0	0.2	1	0.8	0.6	0.23
C_6	1	0.6	0.8	0.11	0.2	0	0.55	0.6	0.2	0.23
C_7	1	1	0.55	1	1	0.55	0	0.11	0.11	0.42
C_8	0.55	0.6	0.2	1	0.8	0.6	0.11	0	0.2	0.23
C_9	1	0.8	0.6	0.55	0.6	0.2	0.11	0.2	0	0.23
C_{10}	0.42	0.22	0.22	0.42	0.22	0.22	0.42	0.22	0.22	0

In study II, we used following distance measure.

$$(c_1, c_2) = \sqrt{(w_{11} - w_{21})^2 + (w_{12} - w_{22})^2 + (w_{13} - w_{23})^2}$$

Thus, the minimal distance “0” occurs when the doctor diagnoses the same *dosha* the first and second time. The maximal distance “1.41” is obtained when two classes have none of the basic types (*vata, pitta,* or *kapha)* in common.

For tongue diagnosis, only three diagnosis classes are present. The chosen distances between these diagnostic classes are shown in Table 7.

Table 7: Distances between tongue diagnoses

Tongue coating and classes	**C1**	**C2**	**C3**
No coating: C1	0	0.5	1
Medium coating: C2	0.5	0	0.5
Tongue coating: C3	1	0.5	0

5.3. Hypothesis of Homogeneous Classification

When a doctor classifies a subject several times, there is a chance that he or she may arrive at different diagnoses. Suppose *n* subjects $j = 1, \ldots, n$ are considered and let p_{jc} be the probability that the doctor chooses classification c for patient j . The hypothesis of homogeneous classification is $H_0 : p_{jc} = p_c$, which is to say that the probability of the doctor assigning a classification of c to subject j does not depend on the subject. In other words, under $H_{0,}$ the doctor is essentially performing random diagnoses according to some common characteristics of the various classes.

5.4. Statistic for Measuring Reliability

Cohen's weighted kappa statistic [43] is a standard measure of intra and inter-rater reliability. Given pairs of classifications ($c_{j1,}$ c_{j2}) for participants $j = 1,\ldots,n$ the weighted kappa is

$$\kappa = 1 - \frac{\overline{D}}{\overline{\overline{D}}}$$

Where, $\overline{\mathrm{D}}$ is the observed average distance between pairs of classifications and $\overline{\overline{D}}$ is the expected average distance under the hypothesis H_0 of homogeneous classifications. The maximal value of 1 for κ is obtained if the observed distance is 0, while κ becomes 0 if the observed average distance is equal to the expected value in the case of random classifications. In general, a larger weighted kappa value means better agreement between pairs of classifications.

5.5. Quantification of Reliability and Permutation Test

One way to assess the magnitude of the weighted kappa statistic is to use the LK scale (Landis and Koch's scale) [44]. However, Bakeman [45] argues that it may be misleading to use one common scale for interpreting kappa since the magnitude of kappa not only depends on observer accuracy (and hence, repeatability) but also on the number of classes and the population probabilities of each class. Another approach to quantifying the magnitude of the weighted kappa is to compare the observed weighted kappa with its distribution under the hypothesis H_0 of homogeneous classification. One can then compute a *p*-value, i.e., the probability of getting at least as favorable a weighted kappa as the observed one, by assuming H_0 is true. As a measure of evidence against $H_{0,}$ *p*-values are comparable across different studies. Smaller *p*-values signify better agreement, but there is not a unique relation between a *p*-value and a specific level of observer reliability according to the Landis and Koch scale.

In practice, we compute the *p*-values using random permutations of the observed classifications, exploring the notion that the diagnoses are independent and exchangeable under H_0. More specifically, we repeat the following two steps many times: 1) randomly permute all diagnoses among subjects and number of classification (first or second) 2) compute the weighted kappa statistic for the permuted data set. Finally, the *p*-value is 1 minus the percentage of permuted data sets for which the computed weighted kappa statistic is smaller than the observed one (Figure 5).

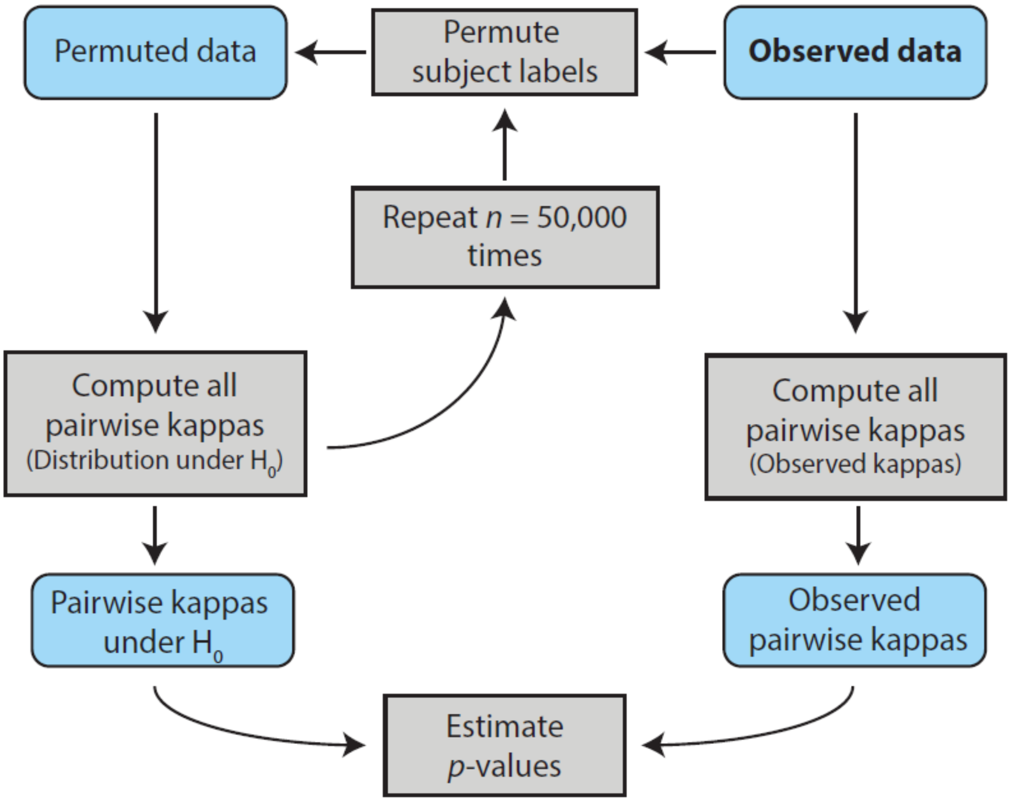

Permuted data
Permute subject labels
Observed data
Repeat *n* = 50,000 times
Compute all pairwise kappas (Distribution under H0)
Compute all pairwise kappas (Observed kappas)
Pairwise kappas under H0
Observed pairwise kappas
Estimate *p*-values

Figure 5: Computing *p*-value

6. RESULTS

Study I evaluated the data obtained by repeated diagnosis for pulse (17) and body constitution (17) by an expert in Ayurveda, with a total of 34+34=68 diagnoses. In study II, 15 doctors diagnosed 20 subjects twice, which generated data for 600 (15×20×2) pulse diagnoses. In study III, 15 doctors examined 20 subjects independently. Thus, data for 300 pulse examinations, 300 body constitution assessments, and 300 tongue diagnoses was collected. Furthermore, self-reporting structured questionnaires and software analyses generated 20 diagnoses each. These data sets were statistically analyzed to quantify the reliability of the diagnostic methods used.

6.1. Results of Body Constitution Assessment

Intra-rater reliability: The hypothesis of homogeneous classification was rejected on the 5% significance level (*p*-values of 0.001). According to the LK scale, the observed value 0.65 of the weighted kappa for body constitution assessment corresponds to “substantial” agreement.

Inter-rater reliability: We computed pairwise kappa for all 105 pairs of doctors. The levels of LK scale for ABC diagnosis were poor (9%), slight (22%), fair (44%), moderate (22%), and substantial (3%) as shown in (Figure 14). The pairwise kappas under permutation are shown in (Figure 11) for only 12 pairs of doctors; the hypothesis of random rating was rejected.

Methods comparison: In study III, we compared a software analysis and a questionnaire assessment with the diagnosis preferred by the majority of doctors for each subject. There was significant evidence against the hypothesis of random rating among software analysis, questionnaire, and the diagnosis preferred by majority of doctors (

Table **8**). Here, a moderate level of inter-rater reliability was present between the most frequent diagnosis versus software assessment, and doctor’s frequent diagnosis versus questionnaire assessment, whereas a fair level of reliability was found between questionnaires and software.

Table 8: The pairwise kappa, *p*-value, and the Landis and Koch scale between the modal diagnosis of all doctors, software diagnosis, and questionnaire diagnosis.

	κ	*p*-value	LK. scale
Modal diagnosis vs. Software	0.487	3e-04	Moderate
Modal diagnosis vs. Questionnaire	0.497	0.0026	Moderate
Software vs. Questionnaire	0.336	0.0128	Fair

✡ - Modal diagnosis: the most frequent diagnosis given to the subject by doctors.

The diagnosis frequency for body constitution: In study I, the doctor frequently diagnosed classes 4, 6, and 5 (*pitta* group). However, in study II, all classes, with the exception of 10, were used, with 2, 5, 6, and 9 (*pitta* combination) being used most frequently and 8, 3, and 4 being used least frequently (Figure 6).

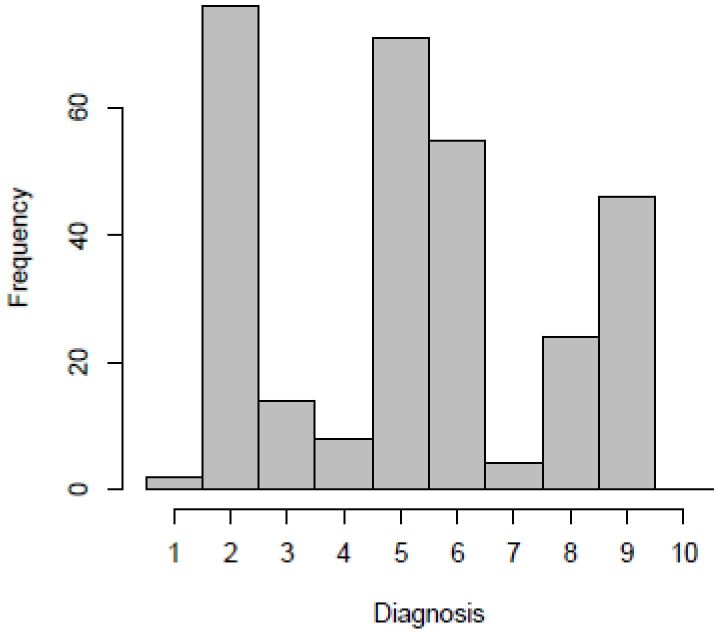


Figure 6: The Frequencies accumulated for all doctors for body constitution assessment for inter-rater reliability.

6.2. Results of Tongue Diagnosis

Inter-rater reliability: We computed pairwise kappa for all 105 pairs of doctors. For tongue diagnosis, levels on the LK scale were poor (16%), slight (37%), fair (41%), and moderate (6%) (Figure 14). The pairwise kappas under permutation are shown in Figure 11. No significant evidence against the null hypothesis was found.

The diagnosis frequency for tongue diagnosis: All three tongue diagnosis classes were used, with class 2 being used most frequently (Figure 7).

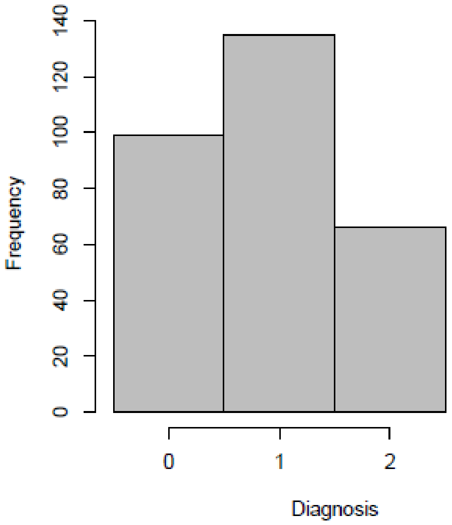


Figure 7: The Frequencies accumulated for all doctors for tongue diagnosis for inter-rater reliability.

6.3. Results of Pulse Examination

Intra-rater reliability: results from study I and study II are shown in Table 9. It showed for each doctor the average distance $\overline{D}$ (i.e. the average of the distances for the 20 pairs of diagnoses made by the doctor), the weighted kappa statistic and a *p*-value for the hypothesis of homogeneous classification (random rating) for each doctor. According to the LK scale, these values of the weighted kappa correspond to low to moderate level of agreement and 2 are even negative.

Table 9: Intra-rater reliability of pulse examination

	Doctor's number	$\overline{D}$ Mean distance	Weighted kappa value	Landis and Koch scale	*p*-values
Study II	1	0.62	-0.18	Poor	0.99
	4	0.55	-0.02	Poor	0.98
	12	0.65	0.03	Slight	0.88
	6	0.21	0.10	Slight	0.82
	2	0.57	0.12	Slight	0.83
	10	0.27	0.15	Slight	0.61
	8	0.38	0.20	Slight	0.58
	14	0.55	0.21	Slight	0.67
	7	0.32	0.26	Fair	0.45
	3	0.35	0.31	Fair	0.47
	9	0.44	0.34	Fair	0.31
	13	0.64	0.36	Fair	0.17
	15	0.36	0.40	Fair	0.11
	5	0.32	0.48	Moderate	0.05
	11	0.28	0.56	Moderate	0.04
StudyI	Single Doctor	0.085	0.42	Moderate	0.02

Inter-rater reliability

In study II, the percentages of pairwise kappas within the LK of categories "poor", "slight", "fair", and "moderate" were 53, 31,15, and 1 percent in the first round, respectively. In the second round, the percentages of pairwise kappas within the LK categories of "poor", "slight", and "fair" were 42, 45, and 13, respectively. The pairwise kappas under permutation are shown in Figure 10. No significant evidence against the null hypothesis was found.

In study III, the levels of the LK scale for pulse examination were poor (40%), slight (37%) fair (20%), and moderate (3%) (Figure 14). The fact that 40 % of pairs had negative values suggests direct disagreement between doctors. The pairwise kappas under permutation are shown in Figure 11. Only 1 pair of doctors performed significantly better than random rating.

The diagnosis frequency for pulse examination: The frequencies accumulated for pulse examination for all doctors were shown in Figure 8 and Figure 9. It shows that all classes except 10 were used, with 2, 5, and 6 being used most frequently and 1, 4, and 9 used least frequently. Out of 15 doctors, 8 doctors mainly use classes 1-6; in addition to these diagnoses 7 doctors also use class 9.

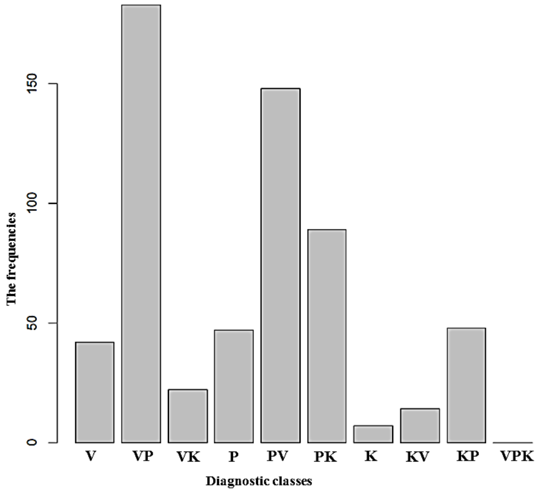


Figure 8: The Frequencies accumulated over all doctors for inter-rater reliability (Study II)

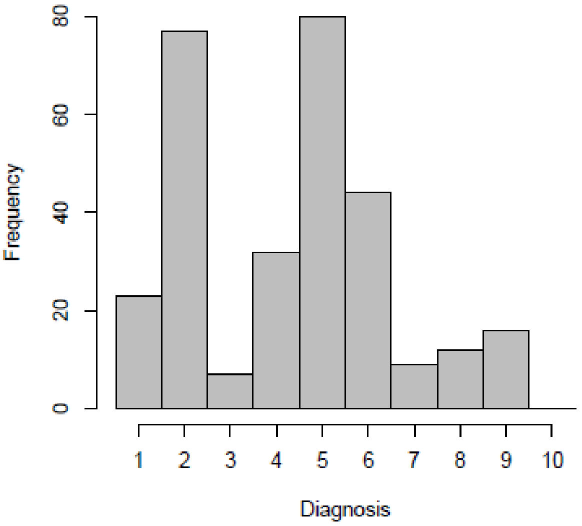
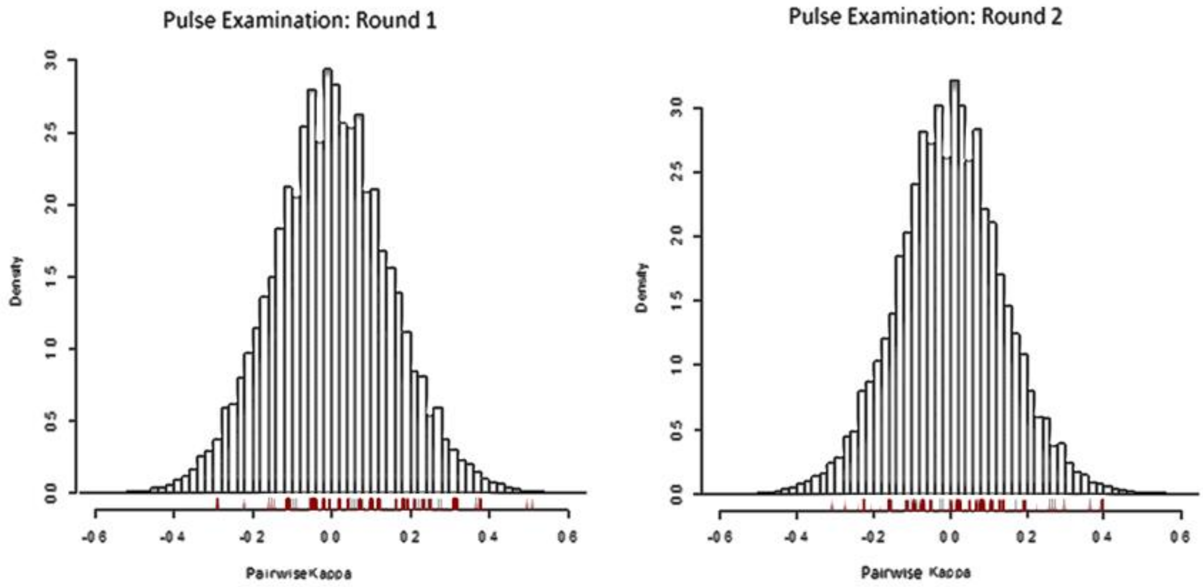


Figure 9: The Frequencies accumulated over all doctors for inter-rater reliability (Study III)



The histogram of all the pairwise kappas under permutation for the two datasets round -1 and round - 2. The red "rug" or ticks below each plot shows the observed 105 pairwise kappas for comparison.

Figure 10: Inter-rater reliability of pulse examination (Study II)

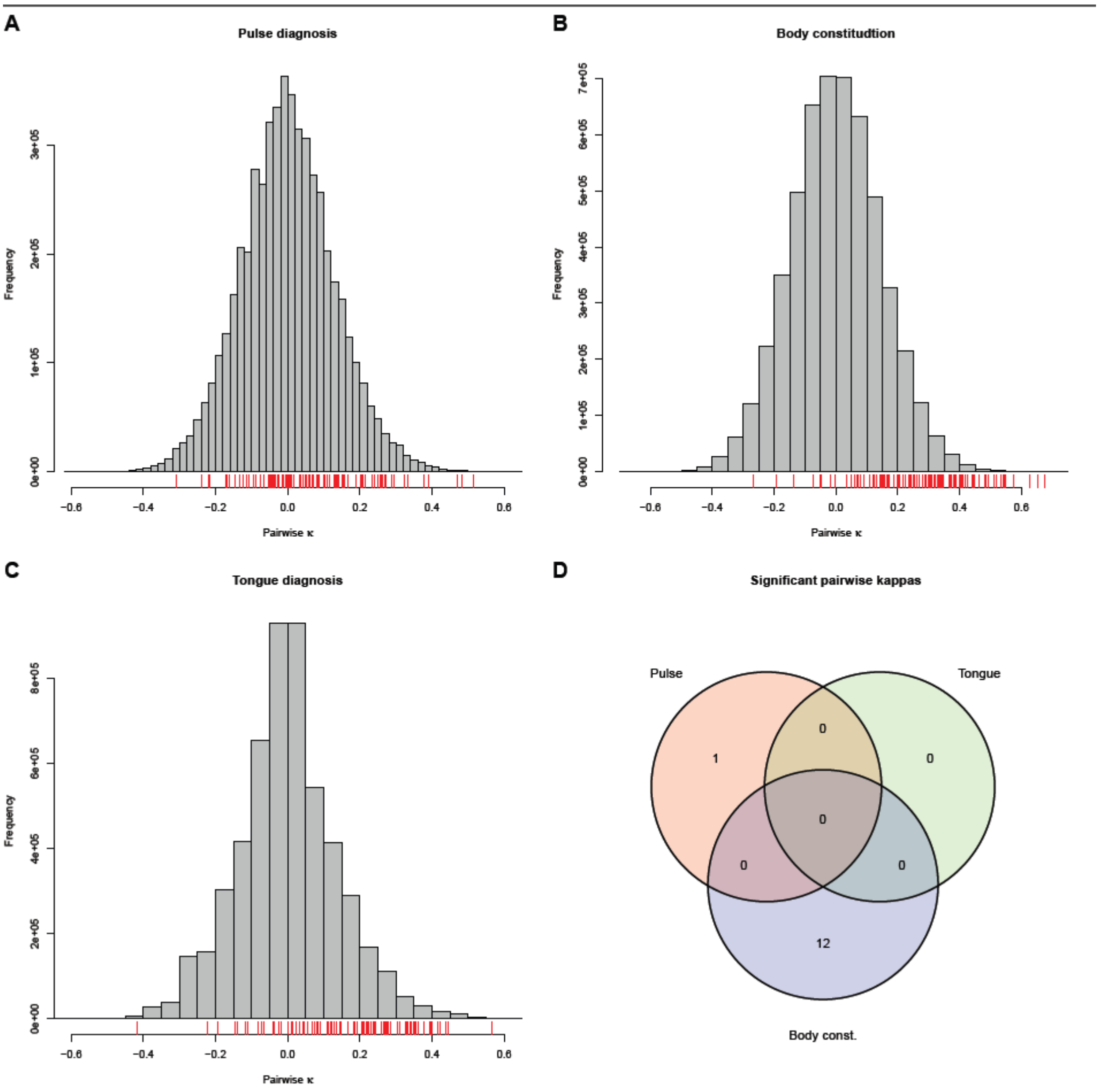


Figure 11: The histogram of all the pairwise kappas under permutation for Pulse (A), Body constitution (B), and tongue (C). The "rug" or ticks in each plot show the observed pairwise kappas for comparison. Panel D shows a Venn diagram of the significant *p*-values in each dataset (Study III).

The relationship between kappa and experience (in number of years) is shown in (Figure 12). The highest kappa was obtained by the doctor with the most experience. Moreover, the kappa values are, in general, higher for the group of experienced doctors and course takers than for the less experienced doctors (*p*-value 0.04 based on one-sided ANOVA) (Figure 13).

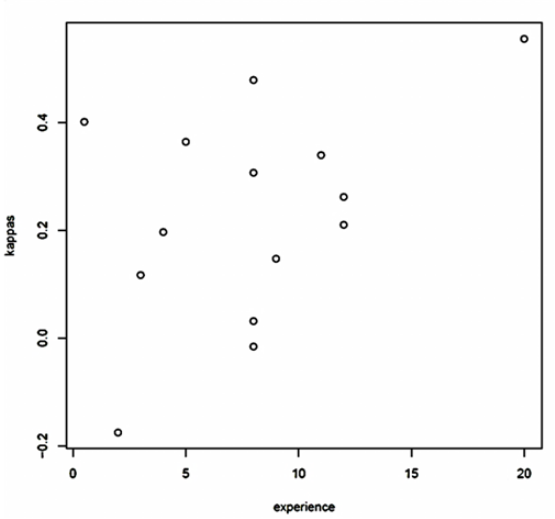


Figure 12: Kappa values vs. experience (in years)

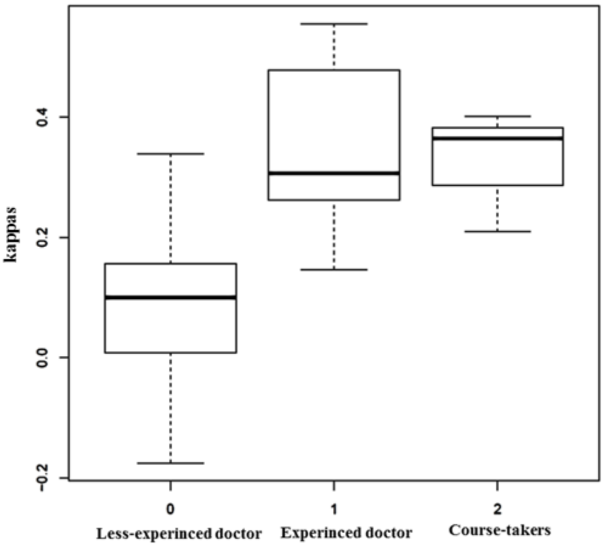


Figure 13: Kappa values for all groups

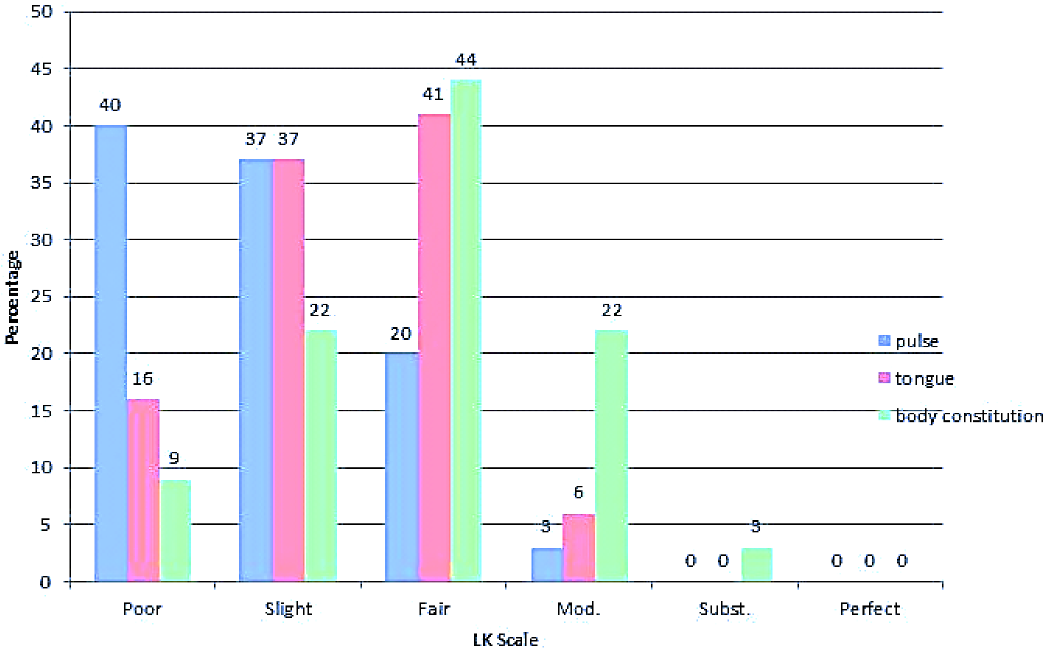


Figure 14: Comparison of pairwise kappas within each LK category (percentage) of reliability for pulse, tongue, and body constitution assessment (Study III).

Table 10 : The average pairwise kappa, the corresponding p-value, and Landis and Koch scale.

Diagnosis	Mean. Kappa	*p*-value	LK. Scale
Prakriti	0.28	2e-05	Fair
Tongue	0.17	2e-05	Slight
Pulse	0.066	2e-05	Slight

The average kappa for *prakriti*, tongue and pulse was 0.28, 0.17, and 0.07 respectively with a corresponding *p*-value less than 2×10-5.

7. DISCUSSION

As the growing need for efficient alternative medicine is felt, research in Ayurveda, as well as other traditional medical sciences is gaining momentum. Today, the experimental, quantitative, and analytical approach of Western medicine is quite relevant and important to Ayurveda. Numerous reliability studies have been conducted in Western medicine but few studies have been done in traditional medicine. This study has assessed the reliability of Ayurvedic diagnostic methods.

7.1. Reliability of Body Constitution Assessment

The reliability of body constitution was estimated by intra-rater (study I) and inter-rater reliability (Study III). The intra-rater reliability was significant, though none of the pairwise kappas were categorized as almost perfect. The hypothesis of random rating was rejected for only 12 pairs of doctors. Body constitution examination includes all three techniques, observation, palpation, and questioning of the subjects, allowing doctors to assess more information than examinations relying on pulse and tongue alone to make a diagnosis. Nevertheless, the results were not promising. Intra-rater reliability was substantial while inter-rater reliability was low, which suggests subjectivity in the diagnosis. However, comparing the self-reporting questionnaires and software analyses with the assessment favored by most doctors did prove significant. This suggested variability between the doctor’s diagnosis and the questionnaire and software analysis. Thus, for reliable body constitution assessment, the doctor’s clinical assessment should be combined with objectively defined parameters such as a questionnaire or software analysis.

Both Sasang Medicine and Traditional Chinese Medicine (TCM), have the same concept of body constitution [46, 47], but very few reliability studies have been conducted on this concept (Table 11). Furthermore, most of these studies studied the validity and internal consistency of questionnaires to assess body constitution. There is only a single study on intra-rater reliability, with a reported range of kappa values from 0.38 to 0.77 based on clinical assessments.

Table 11: Reliability studies conducted on body constitution assessment in traditional medicines

Authors	**Diagnostic method**	**Type of reliability**	**Subject studied**	**Observers**	**Study design**	**Statistical Test**	**Result**
Rastogi, 2012 [33].	PPAT (prototype *prakriti* analysis tool) in Ayurveda	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)
Jang E, 2012 [48].	Sasang Constitution	Intra-rater reliability	86 healthy subjects	Six experts	Experts interviewed subjects twice within an interval of one year.	Cohen kappa	Range for individual expert from 0.380 to 0.768
Ryu H et al, 2010 [49].	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
Yoo J et al, 2007 [24].	Sasangin Diagnosis Questionnaire (SDQ). For assessment of Sasang Constitutional Medicine (SCM)	Intra-rater reliability	88 questionnaires	Self-report structured questionnaire	A questionnaire was administer twice within an interval of two weeks	Pearson's correlation coefficients	0.44 to 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 (κ<0.4)," one item showing "a high degree of concordance," and the remainder (182 items) showing "a moderate degree of concordance"

7.2. Reliability of Tongue Diagnosis

According to LK scale, the overall inter-rater reliability for tongue diagnosis ranged from poor to moderate levels. No evidence against the null hypothesis suggests a low level of inter-rater reliability for tongue diagnosis. In traditional Chinese medicine (TCM) and Korean Medicine, inter-rater reliability for tongue coating showed only a low to moderate level of reliability (Table 12). However, it is reported that an automated tongue diagnosis system (ATDS) showed perfect intra-rater reliability. ATDS was developed to extract information based on a variety of tongue features and then use this to provide doctors with objective information in order to assist in making diagnoses. Due to the assessment of objective features, both intra and inter-rater reliability were strong when compared to the doctors' diagnosis only. In Ayurveda, low levels of reliability for tongue examination could be due to the lack of a standardized tongue examination procedure. The cause of low reliability may also be the lack of specific terminology used to differentiate between thin and thick coating. In TCM, an evidence-based standard was developed to help doctors judge the differences between thin and thick tongue coating [27]. In Ayurveda, future studies and clinical training should utilize precise diagnostic procedures to improve the reliability of tongue diagnosis.

Table 12: Reliability studies conducted on tongue diagnosis in traditional medicines

Authors	Traditional medicine	Type of reliability	Subjects studied	Observers	Study design	Tongue features	Kappa value
Hua B, 2012 [50].	TCM	Inter-rater reliability	40 with knee osteoarthritis	Two TCM practitioners, at least 10 years clinical experience	All patients were seen independently by two experts from the same department at each site.	Coating	-0.04
						Thickness	0.22
Lo L-,2012 [51].	TCM	Intra-rater and inter-rater	20 patients with possible variations in lightning and extruding tongue	12 TCM 3-15 experience	The ATDS is developed to extract a variety of tongue features and provide practitioners with objective information to assist diagnoses. Two sets of tongue images taken one hour apart from 20 patients with possible variations in lighting and extruding tongue.	Intra-rater reliability of ATDS	1.00
						Intra-rater reliability of TCM doctors	0.68 ± 0.27
						Inter-observer agreement between the ATDS and TCM doctors	0.55
						Inter-rater reliability among doctors	0.48
Ko MM, 2012 [52].	Traditional Korean Medicine	Inter-rater	451 subjects with strokes	Two experts well trained in standard operation procedures	All patients were seen independently by two experts from the same department at each site.	Thick coating/fur	0.60
						Thin coating	0.49
O'Brien KA, 2009 [53].	TCM	Inter-rater	45 adults who had mild-to-moderate hypercholesterolemia but who were otherwise healthy.	Three TCM practitioners	Independent assessment		0.22
Ernst E, 2009 [54].	TCM	Inter-rater	55	Two	Field investigation and direct inquiry		0.52

7.3. Reliability of Pulse Examination

The reliability of pulse examination was assessed by intra-rater and inter-rater reliability. The hypothesis of homogeneous classification was rejected on the 5% significance level, which shows that the doctor performed pulse examination in a consistent manner. Study I showed that the distance measure on the categorical pulse and body constitutional diagnosis variables and the permutation test using Cohen's weighted kappa statistic provides a useful statistical methodology for further studies on pulse examination and body constitution assessment.

One may object to the fact that only one Ayurvedic practitioner was used in this study, but the main objective of the study was to investigate the methodology for studying the repeatability of body constitution and pulse examination, i.e., when the same observer repeats the observations. To investigate reproducibility, i.e., inter-rater reliability, we included 15 Ayurvedic doctors in study II and study III.

In study II, each doctor had low kappa values, and some even had negative kappa values. Accordingly, the *p*-values for the hypothesis of homogeneous diagnosis were large (the smallest was 0.04, followed by 0.05, 0.11, and 0.17). This showed a low level of consistency between the two pulse examination sessions for the majority of doctors.

Furthermore, the average inter-rater kappas showed very little agreement between doctors; in fact, the doctors seemed to favor different diagnoses since the proportions of ratings vary. Moreover, study III, the non-blinded study of pulse examination, also showed a low level of inter-rater reliability. The hypothesis of random rating was rejected for the overall test using the average pairwise kappa. According to this the inter-rater agreement can be considered better than for random rating. However, the practical relevance of this can be disputed in light the small average kappa value of only 0.07 and since just one pair-wise kappa was statistically significant separately (Table 10). This variability may be due to the complex qualitative pulse terminologies and the application of different pulse-taking procedures in the practice. Similarly, this low level of agreement is also due to the reliance on different types of traditional evidence, variation in expertise, and the diversity of educational background and expertise of the doctors.

The low reliability of the pulse examination generates debate over whether or not pulse examination should be considered when choosing a health regimen, choice of herbs, or compound formulae. In reality, however, pulse examination is not used as the only deciding factor in these situations; the doctor always confirms his/her diagnosis using other diagnostic methods such as observation, touch, and questioning. In this study, doctors only carried out pulse examination to diagnose *dosha* dominance. This may have influenced the variability.

It was observed that the doctors more frequently diagnosed a combination of two *doshas* c_2 (VP), c_5 (PV), c_6 (PK), than a single *dosha* (Figure 8, and Figure 9). *Pitta* dominates in two of these classes (c_5 & c_6) and is also present in the third (c_2), which indicates that the *pitta dosha* was dominant among the subjects. According to Ayurveda, *kapha*, *pitta*, and *vata* are dominant in children, young, and old people, respectively. In all three studies, most of the subjects were young and displayed *pitta* dominance.

If we compare the results of the reliability studies in other traditional medicine forms in Asia, the majority of studies reported a low to moderate level of reliability, as observed in the present study (Table 13). A comparison between the findings of the reliability studies is difficult because different methods of pulse examination and pulse characteristics were used. Additionally, many studies have not developed formal statistical methods to assess reliability. In this thesis, we have demonstrated results by using a weighted kappa statistic and tested the hypothesis using homogeneous classification.

In both study I, in which the kappa was 0.42, and study II, where it was 0.55, the highest kappa was obtained by the doctor with the most experience. These findings suggest that experience may influence repeatability. It would be of interest to assess the reliability among doctors who have the same number of years of experience. Moreover, in study II, the largest kappa values were obtained by the groups of doctors who had experience in pulse examination or had taken a course. This indicates that practice and proper training can improve repeatability.

Table 13: Reliability studies on pulse examination.

Authors	**Diagnostic method**	**Type of reliability**	**Subject studied**	**Observers**	**Study design**	**Statistical Test**	**Results**
Birch 1997 [21].	Test–retest reliability of pattern diagnosis based on a match between radial pulse examination and abdominal palpation.	Intra-rater	19	One meridian therapy practitioner	Blinded, examination four times in a random order.	Kendall coefficient of concordance	W=0.1053: low but approached statistical significance (*p* = 0.11).
	Test–retest reliability of pattern diagnosis based on pulse diagnosis alone	Intra-rater	35	One meridian therapy practitioner	Blinded examination four times in a random order.		W = 0.0018 (very low).
Kurosu 1969 [25].			40 subjects, half healthy and half with a medical condition	17 meridian therapy practitioners	Blinded examination twice in a random order.	No formal statistical analysis	Healthy subjects 43–65% (mean 53%) Patients: 50–75% (mean 60%). Average percentage = 55.8 %
King E et al, 2002 [28].	Traditional Chinese Medicine, pulse examination	Inter-rater reliability	66 subjects, and in a replication collection, 30 subjects	Two rater	Non-blinded, radial pulse measurement . The initial data were collected on 66 subjects. A replication collection on 30 subjects was completed	Percentage agreement	Averaged 80%

					two months later		
Birch 1999 [21].	Basic pulse qualities in Meridian Therapy	Inter-rater reliability	43	Five	Non-blind, case history was conducted by participating practitioners together then each observer rotated separately through a room.	Kappa	Pulse rate κ=0.29, *p*<0.01 pulse depth κ=0.02, *p*=0.82
O'Brien et al. 2009 [53].			45 patients with hypercholesterolemia who were otherwise health29y	Three TCM practitioners	Non-blinded, independent diagnosis		Pulse location κ = 0.29, pulse force κ = 0.29 (agreement between three practitioners). Agreement between at least two practitioners. Pulse location κ=1.00 Pulse force κ = 0.86 Pulse speed κ= 0.63
O'Brien KA et al, 2009 [29].	Pulse diagnosis in *Toyohari* meridian therapy (TMT)	Inter-rater	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%
Ko MM, 2012 [55].	Traditional Korean Medicine	Inter-rater	658 patients who had strokes	Two clinicians	Non-blinded, independent assessment from February 2010 to December 2010	Kappa value	Kappa range from 0.19 - 0.49 for all pulse types

7.4. Limitations of the Study

The limitation of this study is that there is currently no gold standard to compare the diagnostic accuracy of pulse, tongue, and body constitution. We note that in study I, due to availability of the location and the doctor, trial was conducted in the afternoon. It is generally recommended that pulse examination should be conducted in the morning when the participants are fasting. However, in study II, we excluded this limitation by conducting the examinations in the morning.

Also, different results might have been obtained if we had conducted the pulse examination study on a group of subjects with medical conditions that display well-characterized symptoms as defined in Ayurveda pulse examination. A more heterogeneous group of subjects would allow for more inter-subject variability in the pulse examination and would therefore result in larger weighted kappa values and more testing power for the hypothesis of homogeneous classification.

However, health is defined differently in Ayurveda than it is in the biomedical model. According to Ayurveda, health is defined as the state of equilibrium of bio-entities (*dosha*), digestive juices, enzymes and hormones (*agni*), body tissues (*dhatu*) and the normal excretion of waste materials (*mala*), along with a happy state of soul (*atma*), sensory and motor organs (*indriya*), and mind (*manas*) [56]. This equilibrium tends to be influenced by some unhealthy conditions such as irregular diet, stress, and changes in the weather. As a result, the doctor can be able to diagnose subtle, subclinical changes in the *dosha* in a relatively “healthy” subject.

8. CONCLUSIONS

This is the first study to investigate the reliability of three different and very important diagnostic methods in Ayurveda: body constitution, tongue, and pulse examination. Different types of reliability estimates were used in the study, these being intra-rater reliability, inter-rater reliability, and methods comparison.

In study I, the highest degree of intra-rater reliability (*p*-value 0.001) was obtained with the body constitution assessment, while in study III, the presence of very little evidence against the null hypothesis demonstrated low inter-rater reliability for the body constitution assessment. This essentially demonstrated that while the doctor in study I agreed with his own diagnosis, very little agreement was observed among different doctors in study III, indicating subjectivity in body constitution diagnosis. However, there was significant evidence against random rating by software and questionnaire use, and the diagnosis preferred by the majority of doctors. If we compare the body constitution, tongue, and pulse examinations, doctors performed better in the area of body constitution diagnosis probably due to inclusion of more information for diagnosis, such as observation, palpation, and interrogation.

A lack of evidence against the null hypothesis suggested low inter-rater reliability for tongue diagnosis. The cause of the low reliability may be a lack of specific terminology to differentiate between a thin and a thick coating.

Intra-rater reliability of pulse examination was significant in study I. In study II, intra-rater reliability of pulse examination, according to the LK scale, ranged from a poor to moderate level. The moderate level of agreement observed among the doctors who have more experience and who emphasize pulse examination in their practice indicates that experience may lead to better reliability. Further research is required to establish the reliability of pulse examination by experts in order to assess any effect of training and experience on levels of reliability. Furthermore, assuming homogeneous diagnosis, the doctors used significantly different proportions of ratings, demonstrating the low inter-rater reliability of pulse examination. The low level of reliability potentially reflects inconsistent information about pulse examination procedures and pulse characteristics.

These findings of the reliability studies are not unlike those associated with assessments of the reliability of other traditional forms of medicine, where reliability has also been found to be low. However, this study employed a formal statistical methodology for categorical diagnostic variables. In the statistical analysis, a matrix on the categorical pulse and body constitution diagnostic variables was developed to discuss how the magnitude of the weighted kappa statistic may be interpreted using *p*-values calculated from random permutations of the data. The existing bio-statistical methodology was used in all three studies. Looking ahead, this bio-statistical methodology will be beneficial for future studies of body constitution, tongue, and pulse examination.

The main reason behind the poor reliability of Ayurveda diagnoses could be the lack of a systematic objective methodology and the absence of a precise operational definition of the diagnostic methods. Further studies are required to quantify inter-rater and intra-rater agreement and to gain a greater understanding of the reliability of these diagnoses.

9. IMPLICATION OF THE STUDY AND FUTURE PLANS

On a personal level, acknowledging the variability in the Ayurvedic physical examination is the first step to improving it. For Ayurvedic diagnostic methods to have survived several thousand years to date and developed across many different cultures, there is presumably some observed stability and consistency. Further efforts should be taken to bridge the gap between modern methods and ancient wisdom. The development of a diagnostic guideline based on current scientific evidence is inevitable in contemporary Ayurveda. Thus, it is essential to continue the investigation of the diagnostic techniques used in Ayurveda such as eightfold and tenfold diagnostic methods that contain accepted scientific methods of analysis.

Inter-individual variability in biomedical drug response can be addressed by categorization of subjects. Though, same drug is beneficial for most of the individuals, some of them get adverse reaction. It is possible to minimize or avoid adverse effect by classifying patients. Beforehand, it is necessary to demonstrate reliability of these classification methods.

As observed in the present study, there is a need to understand the reason behind this low reliability. The key to improving reliability is better understanding of the examination technique and its failings and greater standardization of the most effective diagnostic methods. Many studies show that either training of professionals, improving the diagnostic instrument or method, or a combination of both can play a significant role in greater reliability [57]. Furthermore, in TCM, a low level of reliability was observed in two subsequent studies on rheumatoid arthritis ($\kappa = 0.28$ and $\kappa= 0.30$ respectively) [30, 31]. Improvement in the level of reliability ($\kappa= 0.73$) was observed after training sessions for the practitioners from study II [32]. Future studies should be conducted on the benefit of standardizing diagnostic methods and gaining additional training in Ayurveda. Furthermore, these studies should not be limited to diagnostic methods but should also consider treatment plans.

Reliability can be improved by more frequent examinations or having the same patient examined 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 having the same clinician measure blood pressure on several occasions [58].

It is necessary to provide possible objective information of pulse, tongue and body constitution diagnosis for consistent results. Many studied with a technology-focused approach have tried to develop systems that can objectively measure and display the changes in the radial pulse. [59, 60, 61, 62] Automated tongue diagnosis system (ATDS) could yield reliable tongue diagnosis. [27] Further, body constitution analysis can yield reliable diagnosis by incorporating an objectively defined questionnaire and software analysis. Thus, diagnostic methods could be improved by integrating these non-invasive supportive objective techniques.

To demonstrate the reliability of Ayurvedic diagnostic methods, multi-centered, rigorous clinical trials with sufficient subjects and raters must be conducted. Based on the reliability results, clinical reliance should be given to the most reliable variables or methods. In conclusion, the reliability of diagnostic methods is of concern in research, education, and clinical practice. In order for contemporary Ayurveda to be recognized as a credible healthcare system, there is a need for rigorous reliability studies to be performed in the future.

REFERENCES

1. Mishra L-, Singh BB, Dagenais S. Ayurveda: A historical perspective and principles of the traditional healthcare system in India. Altern Ther Health Med. 2001; 7(2):36-42.
2. Mishra L-, Singh BB, Dagenais S. Healthcare and disease management in ayurveda. Altern Ther Health Med. 2001; 7(2):44-50.
3. Sharma H, Chandola HM, Singh G, Basisht G. Utilization of ayurveda in health care: An approach for prevention, health promotion, and treatment of disease. Part 1 - ayurveda, the science of life. Journal of Alternative and Complementary Medicine. 2007; 13(9):1011-9.
4. Sharma H, Chandola HM, Singh G, Basisht G. Utilization of ayurveda in health care: An approach for prevention, health promotion, and treatment of disease. Part 2 - ayurveda in primary health care. Journal of Alternative and Complementary Medicine. 2007; 13(10):1135-50.
5. Khan S, Balick MJ. Therapeutic plants of ayurveda: A review of selected clinical and other studies for 166 species. Journal of Alternative and Complementary Medicine. 2001; 7(5):405-515.
6. Mukherjee PK, Nema NK, Venkatesh P, Debnath PK. Changing scenario for promotion and development of ayurveda - way forward. J Ethnopharmacol. 2012
7. Barnes PM, Powell-Griner E, McFann K, Nahin RL. Complementary and alternative medicine use among adults: United States, 2002. Adv Data. 2004; (343):1-19.
8. Singh RH. Exploring issues in the development of Ayurvedic research methodology. Journal of Ayurveda and Integrative Medicine. 2010; 1(2):91-5.
9. Haynes RB. Of studies, summaries, synopses, and systems: the “4S” evolution of services for finding best current evidence. Evid Based Ment Health 2001; 4(2):37–39.
10. WHO general guidelines for methodologies on research and evaluation of traditional medicine. 2000. Available on : http://apps.who.int/medicinedocs/en/d/Jwhozip42e/
11. Joshi RR. A biostatistical approach to Ayurveda: Quantifying the tridosha. J Altern Comp Med. 2004; 10(5):879-89.
12. Hankey A. The scientific value of Ayurveda. J Altern Comp Med. 2005; 11(2):221-5.
13. 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.
14. Y. T. Acharya, “*Caraka Samhita*,” Chaukhamba Surbharati, Varanasi, India, 1992.
15. Hankey A. Ayurvedic physiology and etiology: Ayurvedo amritanaam. The doshas and their functioning in terms of contemporary biology and physical chemistry. Journal of Alternative and Complementary Medicine. 2001; 7(5):567-74.

16. Abramson JH. Survey methods in community medicine. New York: Churchill Livingstone; 1990.

17. Dunn G. Statistical Evaluation of Measurement Errors: Design and Analysis of Reliability Studies. 2nd ed. London, UK: Arnold; 2004.

18. Streiner DL, Norman GR (1995). Health measurement scales: a practical guide to their development and use. Second edition. 2nd end. Oxford: Oxford University Press.

19. Mulsant BH, Kastango KB, Rosen J, Stone RA, Mazumdar S, Pollock BG. Interrater reliability in clinical trials of depressive disorders. Am J Psychiatry 2002; 159:1598e600.

20. 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.

21. O'Brien KA, Birch S. A review of the reliability of traditional East Asian medicine diagnoses. J Altern Complement Med 2009; 15:353-66.

22. Zaslawski C. Clinical reasoning in traditional Chinese medicine: Implications for clinical research. Clin Acupunct Orient Med 2003; 4:94-101.

23. 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.

24. 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.

25. Kurosu Y. Experimental study on the pulse diagnosis of rokubujoi 11. Jpn Acup Moxib J 1969; 18(3):26–30.

26. 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.

27. 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.

28. King E, Cobbin D, Walsh S, Ryan D. The reliable measurement of radial pulse characteristics. Acupunct Med 2002;20:150-9.

29. 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.

30. 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.

31. 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.

32. 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.

33. Rastogi S. Development and validation of a Prototype Prakriti Analysis Tool (PPAT): Inferences from a pilot study. Ayu 2012; 33:209-18.

34. 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.

35. 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.

36. Hankey A. A test of the systems analysis underlying the scientific theory of ayurveda's tridosha. Journal of Alternative and Complementary Medicine. 2005; 11(3):385-90.

37. S. Rastogi and F. Chiappelli, Bringing Evidence Basis to Decision Making in Complementary and Alternative Medicine (CAM): Prakriti (Constitution) Analysis in Ayurveda, Springer Berlin Heidelberg Publisher, 2010; 7:91-107.

38. S. Rastogi, Evidence-Based Practice in Complementary and Alternative Medicine: Prakriti Analysis in Ayurveda: Envisaging the Need of Better Diagnostic Tools, Springer Berlin Heidelberg Publisher. 2012; 99-111.

39. M. Srinivasulu, "*Concept of Ama in Ayurveda*," Choukhambha Sanskrit Series Office, Varanasi, India, 2010.

40. Lad V. Secrets of the pulse: the ancient art of Ayurvedic pulse diagnosis. Delhi: Motilal Banarasidass Publishers; 2007.

41. Upadhyaya S. Nadi vijnana: ancient pulse science. Delhi: Chaukhamba Sankrit Pratishthan; 2005.

42. Joshi RR. Diagnostics using computational nadi patterns. Math Comput Model. 2005; 41(1):33-47.

43. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull. 1968; 70(4):213-20.

44. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159-74.

45. Bakeman R, McArthur D, Quera V, Robinson BF. Detecting sequential patterns and determining their reliability with fallible observers. Psychol Methods. 1997; 2(4):357-70.

46. 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.

47. 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.

48. 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.

49. 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.

50. Hua B, Abbas E, Hayes A, Ryan P, Nelson L, O'Brien K. Reliability of Chinese medicine diagnostic variables in the examination of patients with osteoarthritis of the knee. Journal of Alternative and Complementary Medicine. 2012; 18(11):1028-37.

51. Lo L-, Chen Y-, Chen W-, Cheng T-, Chiang JY. The study on the agreement between automatic tongue diagnosis system and traditional Chinese medicine practitioners. Evidence-based Complementary and Alternative Medicine. 2012.

52. Ko MM, Lee JA, Kang B-, Park T-, Lee J, Lee MS. Interobserver reliability of tongue diagnosis using traditional korean medicine for stroke patients. Evidence-based Complementary and Alternative Medicine. 2012.

53. O'Brien KA, Abbas E, Zhang J, Guo Z-, Luo R, Bensoussan A, Komesaroff PA. Understanding the reliability of diagnostic variables in a Chinese medicine examination. Journal of Alternative and Complementary Medicine. 2009; 15(7):727-34.

54. Ernst E. Traditional Chinese medicine: How reliable is the tongue diagnosis? MMW-Fortschritte der Medizin. 2009; 151(5):23.

55. Ko MM, Park T-, Lee JA, Choi T-, Kang B-, Lee MS. Interobserver reliability of pulse diagnosis using traditional korean medicine for stroke patients. Journal of Alternative and Complementary Medicine. 2013; 19(1):29-34.

56. Valiathan MS. The legacy of Susruta. New Delhi: Orient Longman; 2007.

57. 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(4):887-95.

58. 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

59. Jog A, Joshi A, Chandran S, Madabhushi A. Classifying ayurvedic pulse signals via consensus locally linear embedding. In: BIOSIGNALS 2009 - Proceedings of the 2nd International Conference on Bio-Inspired Systems and Signal Processing; 20092009. p. 388-95

60. Arunkumar N, Jayalalitha S, Dinesh S, Venugopal A, Sekar D. Sample entropy based ayurvedic pulse diagnosis for diabetics. In: IEEE-International Conference on Advances in Engineering, Science and Management, ICAESM-2012; 20122012. p. 61-2.

61. Begum MS, Poonguzhali R. Noi Kanippaan: Nadi diagnosing system. In: International Conference on Recent Trends in Information Technology, ICRTIT 2011; 20112011. p. 1049-54.

62. Joshi AB, Kalange AE, Bodas D, Gangal SA. Simulations of piezoelectric pressure sensor for radial artery pulse measurement. Materials Science and Engineering B: Solid-State Materials for Advanced Technology. 2010; 168(1):250-3.

APPENDIX

Interpretation of *Sanskrit* words

- *Dosha* - fundamental energies or entities or principles which govern the functions of the body on the physical and psychological level. The Ayurvedic concepts of physiology, pathology, diagnosis, medicine, and therapeutics are based on the doctrine of *tridoshas*.
- *Vata* – a combination of air and ether elements representing kinetic energy and movement, physical or mental functions, and degeneration.
- *Pitta* – a combination of fire and water elements representing thermal energy and metabolism conversion, vision, and emotions.
- *Kapha* – a combination of earth and water elements representing potential energy and structure in the body. It is associated with processes of generation, reunion, and synthesis.
- *Dhatu* - "Dha" means to hold together, to build together"; *dhatu* is the structural, building, element tissue; there are seven tissues defined in Ayurveda: *rasa* (plasma), *rakta* (blood tissue), *mamsa* (muscle tissue), *meda* (adipose tissue), *asthi* (bone marrow), *majja* (bones and nerves), *shukra* (male reproductive tissue) and *artava* (female reproductive tissue).
- *Srotas* - There are innumerable channels in the body; every channel has a root, a passage, and an opening. Within each channel, the physiological function of the respective organ or system is performed.
- *Agni* - The fire element in the body that regulates body heat and performs digestion, absorption, and assimilation of food stuff. It transforms food into energy or consciousness.
- *Ama* - "immature" or "incompletely digested." It is a toxic, unctuous, heavy, and sticky juice which originates as a waste product of digestion and metabolism. *Ama* forms in individuals whose digestion is either weak or overloaded with the wrong foods.