Improving Diagnostic Reliability in Chinese Medicine

Michael C. Popplewell

This thesis is submitted for the partial requirement for the degree of Doctor of Philosophy (Science)

Faculty of Science

University of Technology Sydney

December 2015

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

…….

Signature of Candidate

Abstract

The aim of this thesis was to assess current levels of inter-rater diagnostic agreement in the Chinese Medical (CM) profession and to propose strategies that might improve these levels.

Researchers have generally used inappropriate statistical constructs to evaluate inter-rater agreement. A more appropriate weighted chance-removed statistic is employed to determine inter-rater diagnostic agreement with ordinal data. Further, the largest number of raters which have been used in any past study was three. Similarly, no study was located which involved inter-rater diagnostic agreement with subjects drawn from an open population. This is a deficiency in understanding CM inter-rater agreement in a clinical setting.

The Diagnostic System of Oriental Medicine (DSOM) format was identified as suitable for use in CM diagnosis by practitioners. This format also enables appropriate statistics to be employed. An experiment was performed in which five experienced practitioners of CM diagnosed 42 subjects using the DSOM as the diagnostic format. Each of the sixteen diagnostic descriptors used to describe a diagnosis with the DSOM were scored 0-5. Substantial chance-removed weighted agreement of 0.60 ±0.02 was found. The descriptors of DSOM format were edited after examining 60,000 clinical records at the UTS CM outpatient clinic to arrive at the Chinese Medicine Diagnostic Descriptor

format, the (CMDD). Conventional CM diagnostic formats can be directly mapped to CMDD, thereby making this system as subtle as conventional systems.

A second experiment was performed to evaluate inter-rater agreement with CMDD and contemporary CM diagnostic formats respectively. Groups of CM practitioners, one group utilising the CMDD and the other, the CM diagnostic formats, diagnosed 35 subjects over two days. Each of the fifteen CMDD diagnostic descriptors was scored 0-5, while three selected CM patterns were scored 1-5. The subjects were again drawn from an open population. A weighted simple agreement of only 19% was found between practitioners who employed the CM format. This is not an appropriate foundation for application or assessment of treatment. Further, chance-removed statistics or error estimates cannot be evaluated when the CM format is used with unrestricted diagnostic possibilities.

The possibility that bias was present in raters' scores was also investigated. No significant bias was present in the raters' scores. This should be used as a guide for the adoption of appropriate rater training to improve agreements. Guiding questionnaires for each descriptor whilst utilising the CMDD format, would also appear to hold potential to further improve agreement. The CMDD seems to clearly facilitate superior inter-rater agreement compared with the CM format.

The raters using the CMDD format achieved substantial chance-removed agreements of 0.67 ±0.03 on both days. Mapping diagnoses made by raters

in the CM to the CMDD format enabled chance-removed inter-rater agreements of 0.65 ±0.03 on day one and 0.73 ±0.03 on day two to be calculated, significantly larger than when using the CM format. This suggests that the structure of the CMDD allows the correct inter-rater agreement to be calculated, something very difficult to achieve with the contemporary CM format. It is therefore suggested that the CMDD format be used in contemporary clinical and research settings and is also proposed that it be incorporated into the internationally recognised CONSORT and STRICTA research guidelines

Acknowledgements

I would like to take this opportunity to thank all the people who supported me, especially those who gave me support for this project.

As far as mentors go, I will always be indebted to the late Doctor Phillip Groves, who has had the greatest of influences on me. I was extremely fortunate to meet this incredible man. He showed through his own life that one could make a difference and achieve an incredible amount in a lifetime and I am humbled to have made such minor efforts in comparison to this great man.

I would like to give Associate Professor Chris Zaslawski my sincere thanks for his patience, support and understanding. The project unfolded and often changed directions in ways that all involved would not have ever predicted.

I am extremely grateful to Professor Inseon Lee for providing and processing the Diagnostic System of Oriental Medicine questionnaires, without which this work would have been much more difficult.

I would also like to thank Associate Professor Peter Meier for his generous support in supplying data from the UTS outpatient clinic, which greatly assisted this project.

I would above all like to thank Professor John Reizes, without whose friendship, input, guidance and incredible support this project would not have been possible. John is more like a father than a supervisor to me and it is a priveledge to work with and be educated by this wonderful, generous man.

Table of Contents

List of Figures

List of Tables

Nomenclature

A_0	Number of agreements between two raters observed in an experiment (Equation (2.1));
A_p	Number of possible agreements between two raters (Equation (2.1));
CI	Confidence interval (equation (2.11))
d	Score difference (Equation (2/8))
$\boldsymbol{F}$	Normalising factor (Equations (10.4, 10.8))
$\boldsymbol{F'}$	Rounded normalising score (Equation (10.9))
J	Number of subjects (Equation (10.2))
K	Number of practitioners (equation (10.3))
k	Number of correct answers
L	Number of categories for assessment; (Equation (2.4))
m	m^{th} answer to multiple answer question
M	Number of non-zero scores (Equation (10.8)
N	Number of raters per subject (Equation 2.4))
n	Number of possible ratings (Equation (2.4)
n_o	Number of observations (equation (2.10))
$\bar{\boldsymbol{P}}$	Average simple agreement for more than two raters (equation (2.4))
$\bar{\boldsymbol{P}}_{\boldsymbol{e}}$	Average agreement by chance for more than two raters (Equation (2.5))
$\boldsymbol{p_0}$	Simple agreement between two raters (Equation (2.1))

p_e Agreement by chance for two raters (Equation (2.1))

SE Standard error (equation (2.10))

s Score (equation (10.1))

T Total score (Equation (10.6)

W Weighting

X Difference between average simple agreement and average agreement by chance (Equation (2.15))

Y Difference between perfect agreement and agreement by chance (Equation (2.16))

Greek Symbols

α Ratio of β to the average simple agreement (Equation (2.17))

β Generic definition of agreement with agreement by chance removed (Equation (2.14))

Δ Difference between average simple agreement and the κ_{Fle} statistic (Equation (2.12))

Δ_1 Difference between AC1 and κ_{Fle} statistics (Equation 2.13))

κ $\kappa = \frac{p_0 - p_e}{1 - p_e}$ for two raters or $\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$ for more than two raters

Superscript

$\bar{}$ Mean value

Subscripts

Coh	Cohen
Fle	Fleiss
i, j, k	Indices
l	Linear
q	Quadratic

Glossary and Definition of Terms

Chance-removed agreement (Chapter 2, p. 38): inter-rater agreement from which chance has been removed. For instance if two people were attempting to predict the outcome of a coin toss, there would be a 50% chance that they agreed. The removal of chance agreement is aimed at estimating the 'true agreement' between raters that is not inflated by the presence of chance agreement.

Chinese Medicine (CM): the contemporary style of classic Chinese medicine that was developed in China is practiced in China and Australia. There are other styles of Chinese medicine; Japanese and Korean acupuncture are two that are used by significant numbers of practitioners.

Diagnostic Agreement: diagnostic agreement between practitioners. Diagnostic agreement relies upon the exactly the same terms being used in each practitioner's diagnosis.

Descriptors (Chapter 6 p. 129): are defined as the terms that have specific Chinese medical meaning and form part of the nomenclature of Chinese diagnosis. The Descriptors are the key CM diagnostic attributes used to define Chinese Medical Diagnoses in the DSOM and CMDD. The Descriptors utilised within the CMDD are referenced to the World Health Organisation's International Standard Terminologies of Traditional Medicine in the Western Pacific Region.

DSOM (Chapter 2 p. 70): the Diagnostic System of Oriental Medicine developed by Inseon Lee of South Korea.

DSOMf (Chapter 2 p. 74): The format of the DSOM used for the presentation of diagnostic results of a subject or patient.

DSOM Questionnaire (Chapter 2 p. 72): the diagnostic questionnaire filled in by patients or subjects and used to determine the Descriptors scores of the DSOMf and other data.

Intra-Descriptor Wellness Groups (Chapter 4 p. 104) are defined as groups of subjects allocated to a wellness cohort within a Descriptor according to the scoring characterized by the Total Descriptor Score of the diagnosing raters for that Descriptor.

Simple Agreement (Chapter 2 p 38) is defined as agreement calculated as the number of agreements divided by the number of agreements possible. It is the basis upon all other more complex agreement calculations are made.

Total Descriptor Score (TDS) (Chapter 2 p 38) is the total score allocated by all practitioners to a Descriptor for a particular subject. TDS is used to form the Intra-Descriptor Wellness Groups.

Total Patient Pathogenic Score (TPS) (Chapter 4 p. 116) is defined as the total of all scores allocated to a subject by the diagnosing raters. TPS are used to define Wellness groups.

Wellness Groups are defined as groups of subjects included according to the characteristics of their TPS.

Chapter 1 Introduction

1.1 Oriental Medicine's Current Situation

Chinese Medicine is at a critical point in its long history. Large-scale research projects are being undertaken and the guidelines for such investigations developed. Both endeavors are looking to determine which CM treatments are effective[1-6] in contemporary clinical settings. The ultimate question some researchers are looking to answer is whether there is any provable basis to the historically developed Chinese Medicine theories of diagnosing and subsequent treatment recommendations of textbooks, teaching syllabuses within universities and peers[7].

Already published results can be interpreted to indicate that the current diagnostic models and treatment[6] need a thorough re-examination. At the recent Society of Acupuncture Research conference in Beijing in 2014, the prize for best non-invited speaker presentation was awarded to Ots[8]. After re-analysing verum acupuncture treatment and sham acupuncture[1] data from German acupuncture studies investigating back pain, Ots proposed that acupuncture essentially relied upon dermatome effects and the points and channels were fictional. In Germany, only Western medical doctors and registered alternative health practitioners called heliopracktors are allowed to practice CM and the costs are reimbursed through their public health system. Large-scale randomised controlled trials[9] have been undertaken in Germany, and the efficacy of acupuncture for the treatment of low back pain, knee

[1] Sham acupuncture is a form of placebo treatment where needles are either inserted shallowly or not at all.

osteoarthritis, migraine prophylaxis and tension type headache have been examined. Since only low back pain[10] and knee osteoarthritis[11] were found to improve with treatment with acupuncture, only these two conditions currently attract insurance benefits. Ots asserts that as a consequence, that now in Germany when a medical practitioner gives a patient an acupuncture treatment, it is generally claimed that the patient has been treated for 'back pain'. If Ots assertions were to be interpreted by the German Medical Authorities as indicating that acupuncture is ineffective, CM may not be allowed to be practised in its current form in Germany. If Medical Authorities in other countries interpret Ots results in the same manner the survival of CM as an alternative medical practise may be also be difficult in other countries. If CM moves in the same direction in countries other than Germany, a development towards this direction may take place in other countries as well.

1.2 Statement of problem

Whilst it may seem to be a truism it should be emphasised that without agreement on diagnoses the results of any treatment cannot be evaluated. A major stumbling block in investigating effectiveness of any treatment is the reliability of a diagnosis provided by a CM practitioner. The diversity of diagnostic options and theoretical frameworks available to practitioners makes it difficult to investigate the CM modality. Unlike Western Medicine, which often relies on its many devices and technology for measuring physiology and diagnostic laboratory tests, CM practitioners must rely primarily on the interpretation of verbal information provided by patients. In instances where medical practitioners also have to rely on subjective diagnostic methods, they

also have additional objective data such as x-rays or other types of scans or tests to support their decision-making process. Interestingly in situations where this occurs their levels of agreement are also less than adequate[12-17], suggesting this problem is not isolated to the CM profession exclusively.

Thus, different practitioners may interpret the same "facts" provided by a patient as indicating different health problems. Alternatively two practitioners may interpret disparate facts provided by two different patients as leading to the same diagnosis. This may be a significant contributing factor to the proposition made by some researchers that there is no difference between treatment approaches[6].

Contemporary CM diagnostic models have over one hundred patterns available for selection[18, 19]. Generally two, three or even more patterns are selected. The probability of exact agreement between practitioners is significantly reduced with this extensive framework, particularly as there may be little difference between some patterns.

Previously published inter-rater agreement studies are limited to one condition and a small number of preselected diagnostic options[20-23], or many diagnostic sub-units[24-28]. Thus the level of agreement of diagnoses by multiple practitioners of patients drawn from an open population is unknown; that is, without a pre-diagnosed CM or Western medical condition or defined health status. This is not an appropriate basis upon which we should be investigating CM. Figure 1.2.1 below shows a progression of events that should take place during research into treatment effectiveness.

Figure 1.2.1 Chinese Medicine treatment effectiveness flowchart

Reliable recording of a diagnosis in CM is critical to determining which interventions are effective and which are not. Other influences, such as medications taken for pain or inflammation, anti-depressants, work or home life pressures complicate the patient picture and need to be recorded and considered.

In clinical research and private practice, CM diagnostic reliability is unknown and therefore treatment interventions are possibly not being applied consistently or analysed effectively. Surprisingly, the fact that consistency of diagnosis is unidentified, has not been addressed by panels that formulate research guidelines[29, 30] and thus remains a shortcoming in investigations that are designed to investigate CM treatment efficiency or efficacy.

Objective phenomena related to mechanism of actions of interventions, or the CM system in general should also be re-examined from the perspective of the subject's CM constitutional attributes. Variations in these objective measurements could be correlated to various CM constitutional factors, leading to objective diagnostic tests.

No investigation of levels of diagnostic consensus in the CM profession has been found which was performed on an open population, defined as non-diagnostically categorised subjects. Investigations of pre-diagnosed conditions, as have often been reported[23, 24, 31], do not reveal worthwhile data as to what inter-rater diagnostic repeatability typically takes place in 'real world' clinical settings. Neither does it appear that investigations on open populations have been implemented in any other health modality. This represents a huge deficit in our understanding of diagnostic agreement, and is an impediment to the ongoing development of improved understanding of treatment effectiveness and the mechanisms of treatment actions.

Of equal importance, there seems to be a proliferation of statistical measures employed that are vague, sometimes flawed or inappropriate for determining and interpreting levels of agreement.

1.3 Aims and objectives

Since it has been shown that the statistics generally used do not adequately measure agreement between raters, it is necessary to find a statistical approach to properly calculate inter-rater agreement. This is the first aim of the present work.

It has also been shown that the large number of choices available to practitioners in contemporary CM diagnostic formats leads to a difficult task of determining agreement. It is therefore necessary to develop a diagnostic format, which allows the same flexibility of diagnosis as the contemporary CM

diagnostic method while at the same time allowing the appropriate statistics to be used.

Experiments evaluating inter-rater agreement within open populations when the developed format and the existing CM format are used need to be performed using the appropriate statistics. These experiments will allow the assessment of the suitability and effectiveness of the proposed format for general research and clinical use.

It is generally considered that all experimental results contain random and systematic errors. The systematic errors are termed ‘bias’ in the present work. The statistical methods and diagnostic format should allow for the evaluation and removal of an individual rater’s bias. This perforce necessitates the use of more than two or three raters, particularly when open subject populations are used.

The aim of this thesis therefore is to identify the correct statistics for measuring inter-rater agreement and apply them to quantify agreement between practitioners diagnosing subjects from an open population using the traditional CM diagnostic format, as well as the proposed novel system of diagnosis. Finally, strategies that may improve agreement are investigated.

1.4 Contents of thesis.

A literature review is presented in chapter 2. The first section examines the current use of statistics used in published studies that evaluate reliability. The

second section critically examines the current levels of diagnostic agreement in CM. It is shown that generally the statistics currently being used to ascertain and understand diagnostic agreement are inappropriate and the level of agreement between practitioners has most probably been understated. The correct statistics for determining inter-rater agreement are identified and an example of its use is given. It is also shown that no data collection from open populations has been performed.

Two data collections using open populations were therefore carried out and are described and analysed in chapters 3 to 9. Chapter 3 outlines the details regarding the first data collection, originally used to compare the diagnoses obtained with a patient questionnaire termed the “Diagnostic System of Oriental Medicine” (DSOM) and those produced by five practitioners. The aim was to show the diagnosis obtained with the DSOM was the same as the diagnoses of five practitioners.

The data collected by the five diagnosing practitioners are presented in the DSOM format in Chapter 4 from which the inter-rater agreement between the practitioners is determined. The results obtained with the DSOM questionnaire are also evaluated and compared with the diagnoses made by the five practitioners.

Data from the UTS outpatient clinic is presented in Chapter 5 and compared with the data generated by the present research.

As a consequence of the experience obtained using the DSOM format, a new format is introduced in Chapter 6, termed the “Chinese Medicine Diagnostic Descriptor” (CMDD).

Chapters 7 to 9 deal with a second data collection, in which the CMDD is validated against the contemporary CM method of recording diagnoses. In particular, the details of the second data collection are outlined in Chapter 7. The level of agreement among practitioners using the CM format is presented in Chapter 8, whilst agreement relating to practitioners using the CMDD format is reported in Chapter 9.

In Chapter 10 the hypothesis of practitioner bias being a factor in lowering agreement levels is tested and the results reported.

In Chapter 11 a conclusion, discussion and future directions conclude the thesis.

Chapter 2 Literature Review

This chapter has four sections. The first section is a brief introduction to the concept of diagnosis in CM. The second section reviews the current literature on the biostatistics associated with reliability in a clinical setting. The third section reviews the literature on reliability undertaken in the area of Chinese medicine and acupuncture and the final fourth section reviews the literature on published diagnostic instruments that have been developed for CM.

2.1 The diagnostic procedure in CM

The process of diagnosis in CM is usually a step-by-step procedure. Typically it consists of a loosely structured interview which includes looking, listening touching and smelling[18, 32, 33]. Pattern identification is made by the practitioner through recourse to a number of diagnostic theories and methods; including, but not limited to; eight principles, six channels, qi, blood and body fluid pattern identification approaches[18, 32, 33], and a diagnosis is given taking into account all the information gathered. The interviewing practitioner collects a plethora of diagnostic data, which are essentially decisions regarding many signs and symptoms, which in this thesis are called “diagnostic sub-units”. These diagnostic sub-units are interpreted, weighted and then assembled within the practitioner’s mind and/or in their clinical notes, where a diagnostic landscape that represents the patient’s health from the CM perspective is populated. The end result of this process may lead to a number of diagnostic patterns, usually from one to four or possibly more being ascribed to the

patient. These are described as primary, secondary, tertiary or even quaternary patterns.

Each diagnostic pattern in CM consists either of a single word expression or a combination of terms, somewhat like a phrase used in ordinary speech. There are over one hundred diagnostic patterns used in CM, with some variations in the accepted patterns appearing in different, but widely recognised CM textbooks[18, 19, 32]. This is a consequence of the fact that in any communication, whether oral or written, the same ideas can be expressed in a myriad of ways. Similarly, substitutions or combinations of CM patterns can be used to describe essentially the same condition of a patient. Herein lies a major problem of determining whether two or more diagnoses are in agreement in the CM context and agreement needs to be appropriately defined.

2.2 The Definition of Agreement

The word “agreement” can have many meanings, for the purpose of the present work agreement it is taken to mean a

> “… *situation in which people have the same opinion*” [34]

“The same opinion” requires some examination. **Exact agreement** will only occur if, and only if, **precisely** the same terms in the same order are used in each diagnosis. Since, as mentioned above, often there are many choices for expressing the same diagnosis, the probability of exact agreement is reduced. It follows therefore, that with the large number of possible combinations, the

possibility of exact statistical agreement is unlikely between numbers of practitioners using the above-mentioned contemporary CM models.

Now, since one would expect a practitioner to be consistent in the combinations of terms to describe their patients' diagnostic patterns when the conditions are very similar, the semantic problem may not be so critical when the practitioner compares diagnoses of two or more of their own patients. On the other hand, comparisons between diagnoses made by different practitioners may well be distorted by small semantic differences. Herein lies one of the great problems in evaluating diagnostic agreement in CM.

In order to alleviate this problem, recently researchers have been working on determining diagnostic agreement between practitioners by limiting the pathological categories to be investigated to for example, strokes[27, 28], hypercholesterolemia[25] or rheumatoid arthritis[22, 31]. Even in such cases, it seems that the statistical constructs used to evaluate the level of agreement is flawed.

Unlike the case with Western Medicine in which a diagnosis can be supported by a large number of laboratory and other examinations, in CM the diagnosis is mainly based upon data collected from an interview with a patient. It follows that a practitioner's experience and bias may affect a diagnosis, thereby making the diagnosis less reliable than it would otherwise be. Further, unless there is consensus between practitioners about what exactly constitutes a particular diagnosis, the effectiveness of any treatment procedure performed

by a number of practitioners cannot be reliably evaluated as it may be applied to wrongly diagnosed patients. This is particularly true when dealing with open populations.

Of course the definition of agreement between raters can be made less strict and possibly some judgment might be needed from the evaluators to determine that two raters who may have used different terms in fact meant the same thing. This judgment would further color the results. Indeed, the evaluation of agreement between raters is therefore left to statistical processes, thereby avoiding further confusion. The statistical methods used to measure inter-rater agreement are defined and discussed below.

2.2.1 Simple Agreement

Historically, the first statistical concept used to evaluate agreement was given by

$$p_0 = \frac{A_0}{A_p}, \qquad (2.1)$$

in which p_0 is hereby called the simple agreement (often expressed as a percentage and commonly called “percentage agreement”), A_0 is the number of agreements observed in the experiment and A_p is the number of possible agreements. Equation (2.1) forms the basis upon which all other estimates of the level of agreement.

At present, simple agreement is the most popular method of representing agreement. In a recent review of twenty-eight papers, O’Brien and Birch[24]

stated that thirteen studies, used either percentage agreement or did not indicate the method employed, which is taken here to mean that simple agreement was used. This represents a large proportion, indeed 46%, of the papers reviewed.

Unfortunately this measure, whilst absolutely correct, does not allow for the fact that agreement could have occurred by chance. Agreement by chance is a well-known phenomenon, and distorts outcomes. For instance, suppose that in a multiple choice examination there are four choices for each question and that the correct answers are randomly distributed between the four choices of the questions. A student, totally ignorant of any aspect of a subject, who answered all questions, could expect to obtain a score of 25% by consistently answering the m^{th} ($1 \leq m \leq 4$) answer for each question. Of course if the candidate were to answer each question completely randomly, the probability of answering k questions correctly then becomes $(0.25)^k$ a significant reduction in their score.

What has been assumed in determining the above situations is that the probability of each of the answers to a question is uniformly distributed over all the question; that is that the first, second, third and fourth answer to each question appear $k/4$ times and that k is divisible by 4. Such data are called a "fixed marginal" data. Should this not have been the case, the number of questions that the student answered correctly would have been different to those proposed above.

Similarly, agreement between raters may occur by sheer chance. The first person to attempt to remove the agreement by chance and find the underlying agreement was Cohen[35]. His statistic, κ_{coh}, called Cohen's Kappa, is a measure of the agreement between two raters classifying N items into M categories and is given by

$$\kappa_{coh} = \frac{p_0 - p_e}{1 - p_e} \tag{2.2}$$

in which p_e is the hypothetical probability of agreement by chance calculated from the observed data.

If there is perfect agreement $A_0 = A_p$ in equation (2.1) so that $p_0 = 1$ and $\kappa_{coh} = 1$ in equation (2.2). On the other hand if there is only agreement by chance $\kappa_{coh} = 0$. If the agreement is less than that which would have occurred by chance $\kappa_{coh} < 0$; that is κ_{coh} can also be negative.

Since there are usually more than two assessors, methods were developed for determining agreement with a greater number of raters. Further, since the most "popular" such method, Fleiss' Kappa, is an extension of Cohen's Kappa, the problems engendered by the assumptions made in their derivations will be discussed after the introduction of Fleiss' Kappa.

2.2.2 Kappa Statistics

Cohen's Kappa, the first of the so-called Kappa statistics, appeared in 1960[35]. Unfortunately, as mentioned above, the process developed by Cohen

could only involve two raters and it took until 1971 for Fleiss[36] to extend Cohen's Kappa to include multiple raters. In those circumstances Fleiss proposed that the agreement, κ_{Fle}, taking into account the agreement by chance be given by

$$\kappa_{Fle} = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}, \tag{2.3}$$

in which $\bar{P}$ is the average simple agreement and $\bar{P}_e$ is the average agreement by chance.

Equation (2.3) is very similar to equation (2.2), indeed equation (2.3) becomes equation (2.2) for two raters. Fleiss proposed that $\bar{P}$ and $\bar{P}_e$ be evaluated from

$$\bar{P} = \frac{1}{Nn(L-1)}\left[\sum_{i=1}^{N}\sum_{l=1}^{L} n_{il}^2 - Nn\right] \tag{2.4}$$

and

$$\bar{P}_e = \frac{1}{L}\sum_{l=1}^{L} p_l^2 \tag{2.5}$$

in which,

$$p_l = \frac{1}{Nn}\sum_{i=1}^{N} n_{il} \tag{2.6}$$

N is the number of subjects, n is the number of possible ratings per subject and L is the number of categories into which assessments can be made and n_{ij} is number of raters that assessed the i^{th} subject to the j^{th} category.

Once again κ_{Fle} should fall in the range

$$1 \geq \kappa_{Fle} \geq 0. \tag{2.7}$$

However, as with Cohen's Kappa, negative values of κ_{Fle} are possible and this is discussed by Fleiss[36, 37] But, since this means that the agreement is less than that which would have been obtained by chance, the interpretation of the significance of a negative κ_{Fle} need to be carefully considered in each case.

In the O'Brien and Birch review[24] mentioned above, Fleiss' Kappa was the second most popular approach after simple agreement and was used in 10 of the 28 studies, or 36% of the papers mentioned by them.

The problem is that Fleiss[36], in deriving the estimate for agreement by chance, equation (2.5), followed Cohen[35] in assuming that the data are uniformly distributed between the available choices. Although both Cohen[35] and Fleiss[36, 37] often mention the marginality of the data in their papers, at no point do they clearly state that they have made the assumption that the data need to be fixed marginal, nor do they unmistakeably justify it.

However, in the discussion in section 2.2.1 above of a student totally ignorant of a subject answering a multiple choice examination, the clear assumption was made in the first case that all choices were equally probable taken over all the questions. However, Cohen[35] and Fleiss[36, 37] made the assumption of fixed marginal data implicitly to derive their estimates of agreement by chance. Consequently, neither author justified the implicit assumption of fixed marginal data. It follows therefore that before using either Cohen's or Fleiss'

Kappa, the data need to be tested to ensure that they are fixed marginal, but neither Cohen nor Fleiss issued such a warning in their papers.

Intuitively, it appears that data would rarely be uniformly distributed between the available choices. It appears that many researchers are not aware that there is a restriction on the type of data to which Fleiss' Kappa can be applied and that its use with inappropriate data often leads to wrong interpretations of the results. For example, none of the papers cited by O'Brien and Birch[24] give any indication that the data was tested as to whether it was uniformly distributed between the various choices available to the raters.

Indeed, intuitively the only way that data with fixed margin attributes can be obtained in research projects involving people, appears to be to contrive a situation prior to commencing any experiments in which subjects had been objectively categorised to ensure such an outcome. Whilst this can be easily done with multiple-choice examination paper, an objective pre-diagnosis would be an essential requirement to enable allocation of a uniform distribution of each diagnostic category. Such a setup may not only be impractical, and it would certainly be undesirable in medical research projects. This is due to the fact that in general the distribution between the various categories should **not be expected to be uniform**. Such data are called "free marginal data" and would be expected to occur with diagnostic data of an open population obtained in a CM or any other medical or paramedical setting. It follows that data should be treated as free marginal unless implicitly they are shown to be fixed marginal.

When equation (2.5) is used with Free Marginal Data, the value of $\bar{P}_e$, the agreement by chance, can be significantly overestimated so that the value of κ_{Fle} is drastically reduced[38-42], indicating poor agreement. It follows that unless the marginality of the data is known, poor agreement might not have been the result of a real lack of consensus, but rather than the application of an inappropriate statistical technique to the evaluation of the data. This distortion of agreement has profound and possibly undesirable effects on the interpretation of research outcomes.

In many papers[21, 28, 43, 44], the difference between the simple agreement and the Fleiss' Kappa is very large, κ_{Fle} being much lower than simple agreement. In many cases $\kappa_{Fle}$was so low that researchers have concluded that there was an agreement after agreement by chance had been removed with Fleiss' Kappa. In other cases, this difference, while still significant, does not reduce the value of κ_{Fle}, to the point that agreement is deemed not to exist. It seems interesting that none of the above authors questioned the reasons for the difference between the simple agreement and κ_{Fle} being so large, and accepted the result without objection.

Now, statistics is a complex mathematical subject and for most clinical researchers it is merely a tool rather than a discipline. Perhaps as a result, researchers, without properly evaluating whether the methodology was appropriate, simply used the same statistical approach as their peers had

done in earlier studies. Unfortunately, this may have led to highly misleading interpretations of their results. This apparently, 'blind leading the blind' approach may have led to Fleiss' Kappa being commonly, but wrongly, used as the 'standard' for evaluating inter-rater agreement. The end result has possibly been a systemic distortion and an undermining of the interpretation of results of inter-rater agreement studies.

The information that Fleiss' Kappa should only be used with fixed marginal data has been well known to statisticians for some time. Whilst it had been mentioned as early as 1981[38], and has repeatedly been the topic of papers from that date[39-42, 45, 46], the problem with Fleiss' Kappa has not yet been generally assimilated and other more appropriate methodologies, adopted by medical and paramedical researchers attempting to evaluate diagnostic agreement. An enquiry into Western medicine diagnostic reliability also shows that Fleiss' Kappa appears to have been the most popular choice[15], leading to the conclusion that reliability is also potentially misunderstood in this modality as well.

2.2.3 The AC1 Statistic

The AC1 statistic although first published in 2002[47], is only just being adopted as a inter-rater statistic for general inter-rater agreement evaluations within published studies[27, 28, 48].

Whilst there are other competing statistics[49] including the recently developed PABAK[41] which can only can be used for two raters, or Randolph's Free

Marginal Kappa[49] which has a very slight uptake and no peer-reviewed journal paper to validate the statistic's equations. Gwet's AC1 seems to be the best option, as it apparently addresses the free marginal data issue.

2.2.4 Weighted Statistics

In most cases a simple answer to a diagnostic question is not just yes or no, so that a scale might be introduced. Such a scale could be in words for example indicating intensity of something by say strong, moderate or weak, however, an ordinal scale is usually easier to use and certainly much easier to interpret and manipulate in further mathematical treatments of the data. The first appearance of a solution to provide agreement under these more realistic criteria was undertaken by Cohen and Fleiss in 1973[37], for fixed marginal data and nearly thirty years later by Gwet in 2002[50], for free marginal data. An ordinal scale is needed to enable the calculation of agreement when the severity of a condition is important. Such an approach is termed weighted statistics.

Typically, a diagnosis option should not only be noted that it is present, as appears to be the case in CM, but its severity is surely also important. For example, this approach allows for the easy determination of degree of change in a patient's condition after an intervention.

Only two investigations of CM diagnostic reliability were found that utilised weighted statistics[26, 51] but in both cases Fleiss' Kappa was the underlying statistic. O'Brien[26] did not disclose the type of weighting used and Lo[51] used

the quadratic approach. No papers were found that reported agreement using Gwet's AC2, the term used to describe weighted AC1. Whilst an ordinal scale is easy to construct, with say a six point scale from one to six with one meaning the rater cannot detect any symptom and five meaning as severe as can be imagined, the meaning and therefore the weighting of two raters giving scores which differ by one, two, three or four and possibly the maximum five points apart needs to be established.

2.2.5 Types of Weighting for Ordinal Data

The type of weighting is a point of debate; in essence there is a large number of possible approaches. Three types of weighting are commonly used and have all been programmed into the AC2 software[52], namely, linear, quadratic or radical. Each of these approaches is shown on the graphs, in Figure 2.2.5.1.

The examples in Figure 2.2.5.1 have been based on a six-point scale. As would be expected in each case shown, the maximum weighting of unity is given if raters allot the same score for a particular diagnostic option.

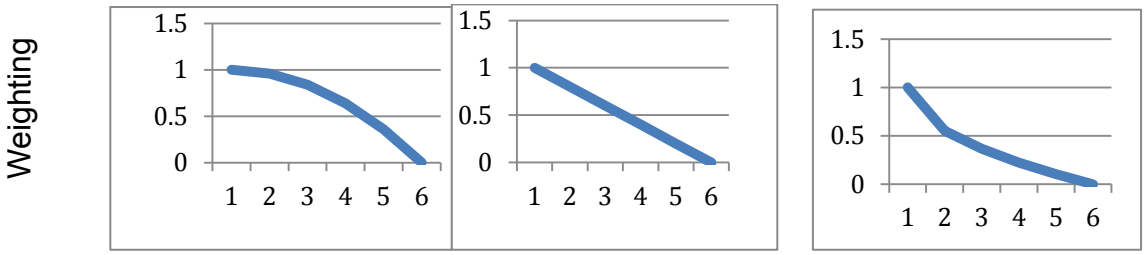

Figure 2.2.5.1 Graph of weightings applied as a function of score difference for (a) Quadratic, (b) Linear and (c) Radical weightings

At the other extreme, if there is a difference of five between two raters, that is one indicating that the diagnostic option is not present and another specifying that the option was present at its maximum intensity, a weighting of zero is assigned in all cases shown in figure 2.2.5.1. The differences between the three approaches deal with how the weights are distributed between these two limits.

In the Linear weighting case the weights are distributed linearly, that is, an increase of one in the point difference reduces the weighting by 0.2, so that the weighting is given by,

$$W_l = 1 - 0.2d,\ 0 \le d \le 5 \qquad (2.8)$$

in which $0 \le d \le 5$ is the weighting, d is the score difference between raters and the subscript l indicates linear.

The quadratically distributed weights used in the software available to the authors, AC2 software[52] is,

$$W_q = 1 - 0.04d^2,\ 0 \le d \le 5 \qquad (2.9)$$

in which the subscript q indicates quadratic

It is not clear how the Radical weighting distribution is obtained so that a table of the weights W_q is given in table 2.2.5.1.

Table 2.2.5.1 Agreement weights applied with radical weighting distribution as a function of score difference

Score Difference	Weight to be applied
0	1
1	0.5527864[2]
2	0.3675445[2]
3	0.2254033[2]
4	0.1055728[2]
5	0

In order to easily interpret equation (2.8 and 2.9) and Table 2.2.5.1, graphs have been plotted in figure 2.2.5.1 to demonstrate in each case the weightings that are attributed to differences between scores given by raters. When Quadratic weighting are used 96% of the full agreement is ascribed if there is a one-point difference between raters. This could be understood to mean that a difference in scores of one unit is so small as to be almost negligible and even a difference in scores of three units is sufficiently small to allocate an agreement of 64%. Supposing that fractions of points can be assigned, it is only when a score difference greater than 3.5 is exceeded that the weighting falls below 50%. This should be compared with when radical weightings are employed when only around 55% agreement is assigned for a difference in scores of one unit. The consequence of the different weightings

[2] The numbers are quoted to the numbers of significant figures generated by the software

is that the highest overall agreement is obtained with quadratic weighting and the lowest occurs with radical weighting.

It can be argued that a difference in score of 20%, represented by a difference of one unit on a six-point scale, is enormous in some circumstances and this is reflected when using the radical weighting approach. In diagnostics, a one-point difference in a six-point scale is surely thought of as reasonably small differences of opinion and this would be reflected by the use of the quadratic weighting scale.

Even after lengthy interviews with a rater, it may not be possible to determine the exact intention of that rater, or what he/she really felt was the difference between say a score of zero and a score of one for a Descriptor. It follows that it is logical to use the linear weighting option, as a compromise between the two extremes of quadratic and radical approaches.

These weightings may seem overly generous in allocating levels of agreement for score selections that have to be considered non-agreement if one rater scores a choice as zero and another as four on a zero to five scale. The linear model for instance gives a 20% agreement in this case. It should be held in mind that the removal of chance agreement would reduce this agreement to what one would rightly expect should be reported.

When using un-weighted statistical approaches, a small difference in opinion between raters indicated by the slightest difference in the allocated score to a

particular diagnostic option is treated as a disagreement. Small differences in opinion as to the severity of a particular diagnostic option indicate that raters have interpreted matters slightly differently, but are generally in agreement. Thus, the increase in agreement obtained when taking weightings into account, can be suggested as indicating that some aspects of practitioner bias had been removed, thereby improving agreement.

Of course if bias exists it is assumed that difference between the raters' scores is consistent. Suppose that the difference in scores between two raters have the same magnitude, but on one occasion one rater gave the higher score, whilst on another case the other rater gave the higher score. Unfortunately, the statistical treatment of the difference is the same in both cases, that is, consistency in the difference between the scores does not influence the results when weights are used. Only the magnitude of the difference between the scores is taken into account, thus the improvement in agreement cannot necessarily be attributed to the use of weighting reducing the effects of bias; although this could be the case. The detection of bias would therefore require other approaches.

2.2.6 Interpretion of Agreement with Agreement by Chance Removed

Kappa and Gwet's AC1 or AC2 statistic results need explanation. While there are other approaches, Landis and Koch[53] suggested the scale shown in Table 2.2.4.1 for interpreting chance-removed inter-rater agreements for

Fleiss' Kappa. It is widely used for this purpose and will be used for interpreting agreement values in the present thesis.

Table 2.2.6.1 Scale for discussing agreements proposed by Landis and Koch

< 0	Poor
0.0 < 0.2	Slight
0.2 < 0.4	Fair
0.4 < 0.6	Moderate
0.6 < 0.8	Substantial
0.8 < 1.0	Almost perfect

2.2.7 Standard Error and Confidence Intervals

Once chance-removed statistics have been calculated, to facilitate proper evaluation of the meaning of the obtained values, Standard Errors and/or Confidence Intervals should be calculated and reported. Simple Agreement is not an estimate; therefore there is little need for measures of variance in this case. Chance Removed Statistics (CRS) such as Kappa, Gwet's AC1 or AC2 are however *estimates* of inter-rater reliability and therefore Standard Errors and the derived Confidence Intervals are therefore necessary.

Standard Errors (SEs) are often confused with the standard deviation of the statistic[54] and are properly called the Standard Error of the Mean value reported, indicating the CRS is the mean value attributable of inter-rater

agreement. The Standard Error of the Means for the CRSs will be always be identified in this thesis as SEs.

The SE of CRS (SE_{crs}) are evaluated with:

$$SE = \sqrt{\frac{p(1-p)}{n_o(1-p_e)^2}} \qquad (2.10)$$

in which n_o is the number of observations by the raters.

Confidence Intervals (CI) are represented by subtracting from the CRS the value of the desired CI level times the SE of the CRS. Given that the most frequent value desired is 95%, the formula uses 1.96 as the constant by which the standard error is multiplied. Thus the confidence limit is

$$\kappa - 1.96SE \leq CI \leq \kappa + 1.96SE \qquad (2.11)$$

Similar expressions are used for other than “Kappa”statistics.

The larger the number of observations measured, the smaller the expected standard error. While a CRS can be calculated for fairly small sample sizes (e.g. 5), the CI for such studies is likely to be quite wide resulting in “no agreement” being within the CI. As a general heuristic, sample sizes should not consist of less than 30 comparisons. According to McHugh[55], sample sizes of 1,000 or more are mathematically most likely to produce a very small CI, which means the estimate of agreement is likely to be very precise.

Interpreting CIs is not straightforward[56]. Most think that CIs are ranges of plausible values for the result or a range of possible values within one

standard deviation. A CI is an estimate of the plausible values for the measure of agreement.

There are many CI misconceptions, as summarised by Kalinowski[56]. The first is the overlap misconception, where comparing the values of two agreements, that the two agreements are statistically significant at $p < 0.5$ when the 95% CIs around the two values are just touching or separate. The CIs are statistically significant at $p < 0.5$ when they overlap by 25%. There is also a confidence level misconception, where it is often believed that a 95% CI for an initial experiment has a 95% chance of capturing the sample mean for a repeat of the experiment. This would be true if the initial sample mean landed directly on the population mean, in fact the average probability is less, around 83%.

Due to the misleading interpretations sometimes given to Confidence Intervals, it was decided within this thesis to report Standard Errors and use the same to determine the difference between CRS values calculated.

2.2.8 Determining differences between two chance-removed inter-rater agreement statistic values

The standard error for the difference between two mean values (here represented as CRS values) is larger than the standard error of either mean and quantifies uncertainty of the differences between the results. The uncertainty of the difference between two means is greater than the uncertainty in either mean. So the SE of the difference is greater than either

SE, but is less than their sum. The SE of the difference between two CRS values is calculated as the square root of the sum of the squares of the two SEs.

2.3 Chinese Medicine Literature on inter-rater agreement

A search of Pubmed, Medline and 'Google' databases was made on the 3rd April 2015. The keywords used were 'inter-rater reliability', 'repeatability', 'TCM', 'Traditional Chinese Medicine', 'TCM diagnostic questionnaires', 'agreement', 'weighted agreement', 'reliability', 'Chinese medicine', 'Kappa', 'Randolph's Free Marginal Kappa' and 'AC2'. Cited studies in each of the papers found from the database search were added to the list and a manual search of the references was also undertaken to identify potential studies.

The criteria for inclusion within this literature review were papers that investigated CM inter-rater diagnostic weighted or un-weighted agreement using simple agreement and also chance-removed Kappa, Randolph or AC2 statistics. The purpose of this criterion was to allow comparison between the Kappa and simple agreement results reported, as the Kappa value on its own could not be trusted as an adequate measure of agreement. If either of these two measures were not specified, the paper was not included in the review. Measures of certainty: standard error or confidence intervals were included in the discussion of each included paper when they were reported, but their absence was not used as exclusion criteria. These were strict requirement indeed for inclusion, as the many studies found did not satisfy these criteria. A review of diagnostic agreement in CM conducted by O'Brien and Birch[24]

included 28 papers, and from this list only seven were suitable for evaluation through inclusion of the specified statistics. Only one study[57] that satisfied these criteria was found outside the list of references given by O'Brien and Birch. Only four of the included papers[21-23, 58] reported measures of certainty in their results and an imposing of the inclusion of measures of certainty would have therefore halved the number of papers reviewed. In total eight papers satisfied the inclusion criteria and were included in the review.

Five papers[20-23, 57], of the eight studied critically reviewed as part of the present study, discussed agreement at the diagnostic pattern level whereas the remaining three[25, 27, 28, 58] examined agreement in many diagnostic sub-groups using un-weighted statistics. All of the papers mentioned above are discussed in section 2.3.1 below. Two of the studies[27, 28] that are part of the nine examined in section 2.3.1, which also utilized the AC1 statistic are further considered in section 2.3.2. In addition in section 2.3.3 two studies[26, 51] are considered that used weighted statistics, with one study[51] employing simple agreement whilst the other did not. This was a digression from the previously stated criteria of having simple agreement to compare with the chance-removed agreement, but due to such limited numbers of papers that satisfied the inclusion criteria an exception was made in this paper.

2.3.1 Chinese Medical Diagnostic Agreement reported with Simple Agreement and Kappa statistics

A summary of the literature on CM diagnostic agreement in which both Kappa and simple agreement were used is presented in Table 2.3.1.1. Simple and

Kappa agreements seems to have been most commonly used statistics for reporting inter-rater agreement in CM for the last decade or so. The range of Kappa values, presented in the second last column, as interpreted by Landis and Koch[53] covered the complete range from 'Poor' to 'Almost Perfect'. The difference between average simple agreement and Kappa values, Δ, is calculated by subtracting Fleiss' Kappa from the average simple agreement values, viz,

$$\Delta = \bar{P} - \kappa_{Fle} \qquad (2.12)$$

Values of Δ were added to the results presented in the original papers in the last column at the right of Table 2.3.1.1. If change were present in terms of a percentage change in the value instead of simply subtracting the two values, the differences would be inflated to very high values when the simple agreement was small. For instance, the largest difference would have been 200% in the case presented by O'Brien *et al*[58] when investigating hypercholesterolemia, within the diagnostic subgroup 'color around eyes'. However, since the actual values are so small, thereby indicating there really is no agreement, the difference between simple agreement and Fleiss' Kappa is meaningless.

Table 2.3.1.1 Chinese Medical diagnostic agreement studies reporting Simple agreement and Kappa values

Researcher	Details	Diagnostic sub-group	Number of Subjects	Number of Practitioners	Number of Options	Simple Agreement	Kappa	Difference between Simple agreement and Kappa Statistic
Macpherson 2004[20]	Back Pain	Primary Diagnosis	12	2	3	0.75	0.00	75
			17	2	3	0.71	0.59	11
			20	2	3	0.60	0.41	19
			15	2	3	0.47	0.02	45
			23	2	3	0.74	0.55	19
		Secondary Diagnosis	8	2	3	0.80	0.67	13
			5	2	3	0.80	0.67	13
			9	2	3	0.56	0.25	31
			11	2	3	0.64	0.44	20
			15	2	3	0.80	0.38	42
Sung 2004[21]	Irritable Bowel	Pre training	39	2	4	0.57	0.11	46
			39	2	4	0.58	0.16	42
		Post training	39	2	4	0.80	0.34	46
			39	2	4	0.81	0.37	54
Zhang 2004[22]	Rheumatoid Arthritis		40	2	10	0.28	0.26	2
			40	2	10	0.26	0.23	3
			40	2	10	0.33	0.30	3
Zhang 2008[23]	Rheumatoid Arthritis	Post training	42	2	10	0.63	0.49	14
			42	2	10	0.86	0.76	10
			42	2	10	0.69	0.53	16

Table 2.3.1.1 Chinese Medical diagnostic agreement studies reporting Simple agreement and Kappa values (continued)

Researcher	Details	Diagnostic sub-group	Number of Subjects	Number of Practitioners	Number of Options	Simple Agreement	Kappa	Difference between Simple agreement and Kappa Statistic
O'Brien 2008[25]	Tongue diagnosis	Constitution	45	3	2	0.48	0.31	17
		Size	45	3	4	0.25	0.20	5
		Color	45	3	4	0.13	0.07	6
		Papillae	45	3	2	0.37	0.16	21
	Tongue diagnosis	Constitution	45	2	2	0.48	0.31	17
		Size	45	2	4	0.90	0.73	17
		Color	45	2	4	0.56	0.19	37
		Papillae	45	2	2	0.37	0.16	21
	Tongue coat	Quality	45	3	5	0.34	0.31	3
		Thickness	45	3	4	0.27	0.22	5
		Color	45	3	5	0.43	0.41	2
	Tongue coat	Quality	45	2	5	0.95	0.90	5
		Thickness	45	2	4	0.95	0.87	8
		Color	45	2	5	0.91	0.9	1
	Pulse Diagnosis	Pulse location	45	3	3	0.24	0.15	9
		Pulse Force	45	3	3	0.37	0.29	8
	Pulse Diagnosis	Pulse location	45	2	3	1.00	1.00	0
		Pulse Force	45	2	3	0.97	0.86	11
		Pulse Speed	45	2	3	0.75	0.63	12
	Spirit	Presence	45	3	2	1.00	1.00	0
		Strength	45	3	3	0.40	0.33	7
	Complexion	Color	45	3	5	0.31	0.28	3

Table 2.3.1.1 Chinese Medical diagnostic agreement studies reporting Simple agreement and Kappa values (continued)

Researcher	Details	Diagnostic sub-group	Number of Subjects	Number of Practitioners	Number of Options	Simple Agreement	Kappa	Difference between Simple agreement and Kappa Statistic
		Color around eyes	45	3	5	0.02	-0.02	4
		Skin texture	45	3	2	0.32	0.09	23
	Hair	Amount	45	3	4	0.51	0.48	3
		Appearance	45	3	2	0.29	0.05	24
	Tongue body	Size	45	3	4	0.25	0.2	23
		Color	45	3	4	0.13	0.07	6
		Constitution	45	3	2	0.48	0.31	17
		Papillae	45	3	2	0.37	0.16	21
	Tongue Coating	Quality	45	3	5	0.34	0.31	3
		Color	45	3	5	0.43	0.41	2
		Thickness	45	3	4	0.27	0.22	5
		Voice Strength	45	3	3	0.60	0.55	5
		Breath character	45	3	3	0.69	0.65	4
	Pulse (right)	Speed	45	3	3	0.75	0.63	12
		Location	45	3	3	0.24	0.15	9
		Force	45	3	3	0.37	0.29	8
	Spirit	Presence	45	2	2	1.00	1.00	0
		Strength	45	2	3	0.90	0.55	35
	Complexion	Color	45	2	5	0.93	0.85	8
		Color around eyes	45	2	5	0.56	0.08	42
		Skin texture	45	2	2	0.32	0.09	23
	Hair	Amount	45	2	4	0.91	0.76	15

Table 2.3.1.1 Chinese Medical diagnostic agreement studies reporting Simple agreement and Kappa values (continued)

Researcher	Details	Diagnostic sub-group	Number of Subjects	Number of Practitioners	Number of Options	Simple Agreement	Kappa	Difference between Simple agreement and Kappa Statistic
		Appearance	45	2	2	0.29	0.05	24
	Tongue body	Size	45	2	4	0.90	0.73	17
		Color	45	2	4	0.56	-0.19	75
		Constitution	45	2	2	0.48	0.31	17
		Papillae	45	2	2	0.37	0.16	21
	Tongue Coating	Quality	45	2	5	0.95	0.9	5
		Color	45	2	5	0.91	0.9	1
		Thickness	45	2	4	0.95	0.87	8
		Voice Strength	45	2	3	1.00	1.00	0
		Breath character	45	2	3	1.00	1.00	0
	Pulse (right)	Speed	45	2	3	0.75	0.63	12
		Location	45	2	3	1.00	1.00	0
Ko 2012[27]	Tongue color	Pale	628	2	2	0.71	0.42	29
		Red	628	2	2	0.76	0.51	25
		Bluish purple	628	2	2	0.92	0.42	50
	Fur color	White fur	628	2	2	0.75	0.49	26
		Yellow fur	628	2	2	0.85	0.69	16
	Fur quality	Thick fur	628	2	2	0.81	0.60	21
		Thin fur	628	2	2	0.75	0.49	26
		Moist fur	628	2	2	0.70	0.29	41
		Dry fur	628	2	2	0.81	0.48	32
	Special appearance	Teeth marked	628	2	2	0.88	0.46	42
		Enlarged	628	2	2	0.90	0.51	39

Table 2.3.1.1 Chinese Medical diagnostic agreement studies reporting Simple agreement and Kappa values (continued)

Researcher	Details	Diagnostic sub-group	Number of Subjects	Number of Practitioners	Number of Options	Simple Agreement	Kappa	Difference between Simple agreement and Kappa Statistic
		Spotted	628	2	2	0.97	0.37	53
	Tongue color	Pale	451	2	2	0.75	0.49	26
		Red	451	2	2	0.77	0.53	24
		Bluish purple	451	2	2	0.94	0.57	37
	Fur color	White fur	451	2	2	0.77	0.52	25
		Yellow fur	451	2	2	0.86	0.71	15
	Fur quality	Thick fur	451	2	2	0.83	0.65	18
		Thin fur	451	2	2	0.75	0.5	25
		Moist fur	451	2	2	0.71	0.31	40
		Dry fur	451	2	2	0.82	0.52	30
	Special appearance	Teeth marked	451	2	2	0.89	0.53	36
		Enlarged	451	2	2	0.91	0.57	34
		Mirror	451	2	2	0.98	0.72	26
		Spotted	451	2	2	0.97	0.4	57
Ko 2013[28]	Pulse location	Floating	658	2	2	0.78	0.36	42
		Sunken	658	2	2	0.83	0.3	53
	Pulse rate	Slow	658	2	2	0.90	0.36	54
		Rapid	658	2	2	0.81	0.46	35
	Pulse force	Strong	658	2	2	0.79	0.47	32
		Weak	658	2	2	0.84	0.49	35
	Pulse shape	String-like	658	2	2	0.78	0.37	41
		Slippery	658	2	2	0.69	0.38	31
		Fine	658	2	2	0.85	0.46	39

Table 2.3.1.1 Chinese Medical diagnostic agreement studies reporting Simple agreement and Kappa values (continued)

Researcher	Details	Diagnostic sub-group	Number of Subjects	Number of Practitioners	Number of Options	Simple Agreement	Kappa	Difference between Simple agreement and Kappa Statistic
		Surging	658	2	2	0.91	0.39	52
	Pulse location	Floating	451	2	2	0.78	0.4	38
		Sunken	451	2	2	0.83	0.34	49
	Pulse rate	Slow	451	2	2	0.92	0.46	46
		Rapid	451	2	2	0.80	0.48	32
	Pulse force	Strong	451	2	2	0.79	0.48	31
		Weak	451	2	2	0.84	0.49	35
	Pulse shape	String-like	451	2	2	0.81	0.43	38
		Slippery	451	2	2	0.70	0.4	30
		Fine	451	2	2	0.84	0.47	37
		Rough	451	2	2	0.94	0.17	77
		Surging	451	2	2	0.92	0.47	45
Grant 2013[57]	Prediabetes		27	2	3	0.70	0.56	14

Firstly, brief comments upon the methodologies, the inter-rater agreement reached and appraisals of the overall weakness or strengths of the reviewed studies will be given next in section 2.3.1.1. Then more generalised observations and insights regarding the results reported in these papers will be made in section 2.3.1.2.

2.3.1.1 Comments on papers that reported agreement with average simple agreement and Fleiss' Kappa

Macpherson *et al's*[20] paper appeared to be focus upon reporting diagnoses selected and treatments provided for back pain than the inter-rater reliability of the practitioners, so little detail regarding the inter-rater reliability was provided. Five qualified practitioners each with minimum of five an average of 12.8 years experience were blinded from and compared to Macpherson, the primary investigator's diagnosis, in varying numbers of cases from 8 to 23. While 148 were included in the study, 87 primary and 48 secondary diagnoses were investigated randomly to confirm inter-rater diagnostic consistency. The Kappas derived varied widely; from 0.00 to 0.59 with a Fair[53] average of 0.31 for the primary diagnoses and from 0.67 to 0.25 with a Moderate[53] average of 0.48 for the secondary pattern were reported. There were large variations in inter-rater reliability in this study.

Sung *et al*[21] reported high variability in diagnoses and corresponding Slight[53] average Kappa agreement of 0.13 before training and a slightly improved Fair[53] 0.35 after training. They believed the results were due to differences in training, education and experience but the improvement in inter-rater

agreement after training still did not attain desirable levels. The quality of the interviewing practitioners utilised appeared sound, with four CM qualified practitioners who were blinded from each other and who reported having more than five years clinical experience, interviewing each of the 39 subjects in pairs.

Zhang *et al*[22] investigating diagnostic and herbal prescription agreement in rheumatoid arthritis patients in 2004 found a Fair[53] average Kappa agreement of 0.26 and again like Sung *et al* pointed to differences in practitioner training or experience. The rating practitioners however each had least five years experience and were all graduates of a five-year herbal medicine program so this should not be a significant factor. Four primary CM patterns were identified that could be ascribed to sufferers of rheumatoid arthritis which were increased to ten upon combinations of each of these four and their combination with lesser seen diagnostic patterns. Using these ten diagnostic options a Fair agreement was found. In an attempt to improve agreement, less stringent criteria for agreement were attempted, where the terms used to describe the patterns were separated. For instance, if any pattern included the term cold or heat, agreement was deemed to have occurred. These criteria produced improved average Kappas of between 0.55 and 0.87. The majority (81%) of subjects were ascribed three or fewer CM patterns. Raters saw each patient in a randomised format called a modified 'Latin Square method' to avoid result contamination from interview sequence factors. In Zhang *etal's*[23] second paper published in 2008, using the same practitioners as the 2004 study, agreement was improved to a Substantial[53]

average Kappa of 0.73 after training, an impressive difference. While it is not absolutely clear, it seems that the more generous second method of determining diagnostic agreement proposed in Ko *et al's* first paper was adopted in Ko *et al's* second paper as the primary approach for valid agreement. If this was the case, the results were similar to the earlier study.

O'Brien *et al's* paper[25] which investigated hypercholesterolemia, looked at many diagnostic sub-units, with between two and five options in each sub-unit, and compared the diagnostic choices of three raters with 45 patients. Two of the raters were trained in China and had over twenty years' clinical experience each and the last one had university training in Australia and five years part-time clinical experience. Two different criteria were used to ascribe agreement: where all three raters agreed or where only two did. Better results were found if agreement between two raters was accepted as the criteria of agreement, with simple agreement of 0.79 and a Moderate[53] Kappa of 0.62 for two raters compared to only 0.38 simple agreement and a Fair[53] Kappa of 0.30 for three raters. Confidence intervals and standard errors were both reported. Where Kappa results were below 0.50, standard errors were often 50% of the Kappa result, accompanied by greatly expanded confidence intervals, both suggesting the Kappa values in these cases were not certain. Where the Kappa results were above 0.50, both measures of certainty were improved.

O'Brien *et al*[58] with the same team also reported upon diagnostic agreement using eight guiding principles with the same subjects and raters, but as the

diagnostic categories were open ended, no valid Kappa statistics could be calculated and therefore the agreement cannot be included in this review.

Ko *et al* in 2012[27] and 2013[28] produced two papers that both investigated stroke, with the same 451 patients in a multisite study, both reporting on many diagnostic sub-units, the first relating to the tongue and the second relating to the pulse diagnosis. There is no mention of the rater's experience other than to say that they were 'experts' in either paper. Average Kappas from all diagnostic sub-units were both Moderate[53]: 0.51 in the first paper and 0.40 in the second. Confidence intervals were an average ±20% in the first study and ±30% in the second. These confidence intervals seem large, even though the number of subjects included was the highest of all papers reviewed.

Grant *et al*[57] in 2013 used a questionnaire called TEAMSI TCM[59] to influence, but not replace the diagnostic choices of two diagnosing practitioners diagnosing 27 Western medically diagnosed pre-diabetic subjects. The use of TEAMSI TCM seemed an interesting approach to try to improve agreement that appeared to be successful; reflected by a higher level of inter-rater agreement than was commonly observed in other similar papers, that of a Moderate[53] Kappa value of 0.56, but with a large 95% confidence interval of 0.25-0.81, likely indicating the sample was possibly too small. The practitioners used each completed four-year Bachelor degrees from the University of Western Sydney and all had more than four years clinical experience.

2.3.1.2 Generalised observations regarding Fleiss' Kappa and average Simple Agreement studies

In this section, firstly the choices available to raters, then numbers of raters utilised, then number of subjects rated will be investigated. How each variable seemed to affect overall agreement will be presented and discussed.

Number of diagnostic options made available

The number of diagnostic options that were made available to raters by the researchers whose work is summarised in Table 2.3.1.1 and listed in column 5 of that table. The number of choices offered to the raters in each sub study is now summarised in table 2.3.1.2.1.

Table 2.3.1.2.1 Number of options made available to raters by researchers using Fleiss' Kappa and simple agreement

Options offered	Number of results	Proportion of total studies
2	62	49%
3	28	22%
4	18	14%
5	12	10%
10	6	5%

In almost half of the studies cited in this literature review, only two options were presented to the raters. On the average, including the two studies published by Zhang *et al*[21, 23] mentioned above, which included ten possible diagnostic outcomes the number of options was 3.2.

The average simple agreement was 0.69 across all studies reviewed. Where there were two options available to the raters, average agreement was 0.75. Where there were three options, the average was 0.71. The average simple agreement when four options were available was 0.58. Five possible selections produced average simple agreement of 0.59 and ten diagnostic options led to mean simple agreement of 0.51. Average simple agreement dropped as the number of options increases, but somewhat plateaued if more than four options were available to the raters.

Average Kappa agreement, due to the removal of the estimated chance-removed agreement, was always lower than simple agreement. The average across all results independent of number of options was 0.45. An interesting observation was with the average Kappa value not reducing in the same predictable, orderly way when examined according to number of rater options. Different choices produced unexpected average Kappas when compared to the average simple agreements reported. Two resulted in 0.42, three 0.62, four 0.37, five 0.52 and ten 0.43. there seemed no logical reason for these changes in average Kappa values.

The differences between simple agreement and Kappa are presented in the last column of table 2.3.1.1. The average differences, as a product of the scores presented in the two columns to its left are again unpredictable due to the large variations in average Kappas. This average difference is greatest within the studies that offer two choices, at 0.33, this average reduces to only

0.09 where three choices are available, jumps up to 0.21 with four choices and again drops to 0.7 and 0.08 in the five- and ten-option studies.

Number of Raters Compared

The robustness of all studies included in the literature review would have benefited from an increase in the number of practitioner diagnoses made to each patient. Only two or three practitioners were compared in all studies cited. Where two practitioners were compared, higher average agreements; average simple agreement of 0.77 and Kappa of 0.49 were reported, compared to only 0.38 and 0.30 where three were compared. Increases in the number of diagnosing raters were therefore associated with decreases of the levels of agreement reported.

Subject Numbers

The number of subjects included in the studies ranged from 8 to 628, with four studies of the eight including approximately 40 subjects[21-23, 58]. Larger numbers of participants increase the possibility of reduced standard errors or increased confidence intervals, which were only reported in half the papers[27, 28, 57, 58]. In many cases when confidence intervals or standard errors were reported by the researchers, they were quite large indicating that the statistical validity of the results is questionable

All these papers used Fleiss' Kappa, so while the comments made by the researchers were appropriate if values derived by the Kappa statistic used was the correct choice, the agreement values reported would often be lower

than what would be derived with the more appropriate statistic, the AC1, indicating that agreement is possibly better than what has been reported.

2.3.2 Chinese Medical inter-rater agreement evaluated with the AC1 statistic

The AC1 statistic is only just being adopted as an inter-rater statistic for general inter-rater agreement determination. Since the approach to the evaluation of the AC1 value addresses the free marginal data issue, it seems to be the best option available at present to overcome the difficulties of Fleiss' Kappa. Two investigations[27, 28] included in table 2.3.1.1 above, also presented diagnostic agreement calculated with the AC1 methodology in addition to the commonly used simple agreement and Fleiss' Kappa. The results of these studies are again summarised in table 2.3.2.1, but this time with the addition of the AC1 data.

Table 2.3.2.1 Papers reporting percentage, Kappa Number and AC1 agreement summarised

Researcher	Details	Diagnostic sub-group	Number of Subjects	Number of Practitioners	Number of Options	Simple Agreement	Kappa	AC1	Difference between AC1 and Kappa Statistic
Ko 2011[27]	Tongue color	Pale	628	2	2	0.71	0.42	0.43	0.01
		Red	628	2	2	0.76	0.51	0.52	0.01
		Bluish purple	628	2	2	0.92	0.42	0.9	0.48
	Fur color	White fur	628	2	2	0.75	0.49	0.51	0.02
		Yellow fur	628	2	2	0.85	0.69	0.71	0.02
	Fur quality	Thick fur	628	2	2	0.81	0.6	0.63	0.03
		Thin fur	628	2	2	0.75	0.49	0.49	0.00
		Moist fur	628	2	2	0.7	0.29	0.49	0.20
		Dry fur	628	2	2	0.81	0.48	0.68	0.20
	Tongue appearance	Teeth marked	628	2	2	0.88	0.46	0.84	0.38
		Enlarged	628	2	2	0.9	0.51	0.86	0.35
		Mirror	628	2	2	0.97	0.6	0.97	0.37
		Spotted	628	2	2	0.97	0.37	0.96	0.59
	Tongue color	Pale	451	2	2	0.75	0.49	0.51	0.02

Table 2.3.2.1 Papers reporting percentage, Kappa Number and AC1 agreement summarised (continued).

Researcher	Details	Diagnostic sub-group	Number of Subjects	Number of Practitioners	Number of Options	Simple Agreement	Kappa	AC1	Difference between AC1 and Kappa Statistic
		Bluish purple	451	2	2	0.94	0.57	0.93	0.36
	Fur color	White fur	451	2	2	0.77	0.52	0.56	0.04
		Yellow fur	451	2	2	0.86	0.71	0.74	0.03
	Fur quality	Thick fur	451	2	2	0.83	0.65	0.68	0.03
		Thin fur	451	2	2	0.75	0.5	0.51	0.01
		Moist fur	451	2	2	0.71	0.31	0.5	0.19
		Dry fur	451	2	2	0.82	0.52	0.72	0.20
	Tongue appearance	Teeth marked	451	2	2	0.89	0.53	0.86	0.33
		Enlarged	451	2	2	0.91	0.57	0.88	0.31
		Mirror	451	2	2	0.98	0.72	0.98	0.26
		Spotted	451	2	2	0.97	0.4	0.97	0.57
Ko 2013[28]	Pulse location	Floating	658	2	2	0.78	0.36	0.66	0.30
		Sunken	658	2	2	0.83	0.3	0.77	0.47
	Pulse rate	Slow	658	2	2	0.9	0.36	0.89	0.53

Table 2.3.2.1 Papers reporting percentage, Kappa Number and AC1 agreement summarised (continued).

Researcher	Details	Diagnostic sub-group	Number of Subjects	Number of Practitioners	Number of Options	Simple Agreement	Kappa	AC1	Difference between AC1 and Kappa Statistic
		Rapid	658	2	2	0.81	0.46	0.71	0.25
	Pulse force	Strong	658	2	2	0.79	0.47	0.65	0.18
		Weak	658	2	2	0.84	0.49	0.77	0.28
	Pulse shape	String-like	658	2	2	0.78	0.37	0.67	0.30
		Slippery	658	2	2	0.69	0.38	0.38	0.00
		Fine	658	2	2	0.85	0.46	0.79	0.33
		Rough	658	2	2	0.93	0.19	0.93	0.74
		Surging	658	2	2	0.91	0.39	0.9	0.51
	Pulse location	Floating	451	2	2	0.78	0.4	0.64	0.24
		Sunken	451	2	2	0.83	0.34	0.79	0.45
	Pulse rate	Slow	451	2	2	0.92	0.46	0.9	0.44
		Rapid	451	2	2	0.8	0.48	0.68	0.20
	Pulse force	Strong	451	2	2	0.79	0.48	0.65	0.17
		Weak	451	2	2	0.84	0.49	0.76	0.27
	Pulse shape	String-like	451	2	2	0.81	0.43	0.72	0.29

Table 2.3.2.1 Papers reporting percentage, Kappa Number and AC1 agreement summarised (continued).

Researcher	Details	Diagnostic sub-group	Number of Subjects	Number of Practitioners	Number of Options	Simple Agreement	Kappa	AC1	Difference between AC1 and Kappa Statistic
		Fine	451	2	2	0.84	0.47	0.77	0.30
		Rough	451	2	2	0.94	0.17	0.94	0.77
		Surging	451	2	2	0.92	0.47	0.91	0.44
Averages				2	2	0.83	0.46	0.72	0.26

These two papers were previously summarised in 2.3.1. Predominantly Substantial[53] and Almost Perfect[53] agreements were reported with AC1, while significantly lower predominantly Moderate[53] agreement was found with the same data with Fleiss' Kappa demonstrating the significantly lower values reported with Fleiss' Kappa when compared to the AC1 statistic.

Δ_1, the difference between the AC1 statistic and Fleiss' Kappa values is calculated by the subtraction from the AC1 of the Kappa value, viz,

$$\Delta_1 = AC1 - \kappa_{Fle}, \qquad (2.13)$$

in which AC1 is the AC1 value. $\Delta_1$is presented in the last column of Table 2.3.2.1 in addition to the data obtained from the papers. On the average difference the AC1 value is 0.26 higher than Fleiss' Kappa, but this does not adequately describe the situation. In some cases the difference is very large; in two cases the disparity is $\Delta_1 > 0.7$and in five of the 48 cases, 10% of all values of Δ_1, it is ($0.7 < \Delta_1 < 0.5$). A further five results, or another 10%, also differed by ($0.5 < \Delta_1 < 0.4$).

In most cases however, the difference was smaller, as in 14 cases of the 48 subgroups, or 29% of all cases, the difference is less than 0.10 and a further seven results had differences of 0.20 or less, a total of 44% of all results investigated by Ko *et al*[27, 28]. They also mention that Fleiss' Kappa has been previously used by many investigators, but that its application may lead to

problems in interpretation[27]. They therefore discuss the inter-rater agreement from the AC1 perspective only. The large variation between the Fleiss' Kappa and AC1 values that sometimes occur further confirm and illustrate the difficulties that may arise using Fleiss' Kappa to evaluate inter-rater agreement.

The simple agreement and AC1 agreement values for each diagnostic sub-unit presented in table 2.3.2.1 ordered from the lowest to the highest and are presented in Figure 2.3.2.1. It may be seen that as simple agreement increases the differences between the two statistics reduce so that they are virtually identical as the simple agreement approaches 1. On the other hand, as simple agreement reduces, the difference between the two statistics increases. Indeed, the divergence between the two increases quite markedly if the simple agreement is less than about 0.8 and at about 0.7 the AC1 values are approximately about half those of the simple agreement.

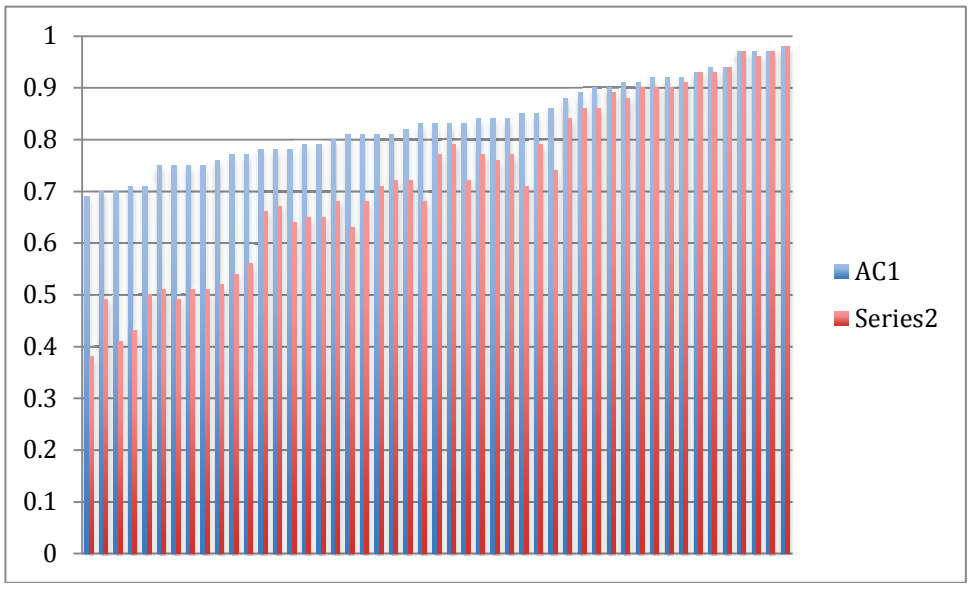

Figure 2.3.2.1 Simple agreement and AC1 statistics from table 2.3.2.1 sorted according to simple agreement

The reasons for this behavior follow from the definition of the AC1 and Fleiss' Kappa statistics. Let β be the generic definition of agreement with agreement by chance removed. β has the form

$$\beta = \frac{X}{Y} \tag{2.14}$$

in which $$X = \bar{P} - \bar{P}_e \tag{2.15}$$

and $$Y = 1 - \bar{P}_e. \tag{2.16}$$

Equation (2.14) is in fact a rewrite of Equation (2.3) so that β represents either AC1 or Fleiss' Kappa depending on how $\bar{P}_e$, the agreement by chance is evaluated. It follows that both Fleiss' Kappa and AC1 would behave in the same way as functions of $\bar{P}$ and $\bar{P}_e$.

When $\bar{P} = 1$, independent of the value of $\bar{P}_e$ $\beta = 1$, thereby indicating perfect agreement. Now suppose that $\bar{P}_e$ remains constant as $\bar{P}$ is varied. As $\bar{P}$ is reduced the numerator in equation (2.14) becomes smaller independent of the value of $\bar{P}_e$. For a particular value of $\bar{P}_e$, the denominator in equation (2.14) remains constant so that β decreases as $\bar{P}$ is decreased. This is clearly seen in Figures 2.3.2.2 and 2.3.2.3.

The gradient of the lines of constant $\bar{P}_e$ in Figure 2.3.2.2 increases as the agreement by chance increases. This means that for a situation in which the simple agreement is, say 0.7 and the agreement by chance is, say 0.2, the

agreement with agreement by chance removed becomes 0.625. However, if the agreement is 0.7 and the agreement by chance is 0.4, the agreement with agreement by chance removed becomes 0.5. Finally if the simple agreement is 0.7 and the agreement by chance is 0.7, as would have been expected, the agreement with agreement by chance removed now becomes zero.

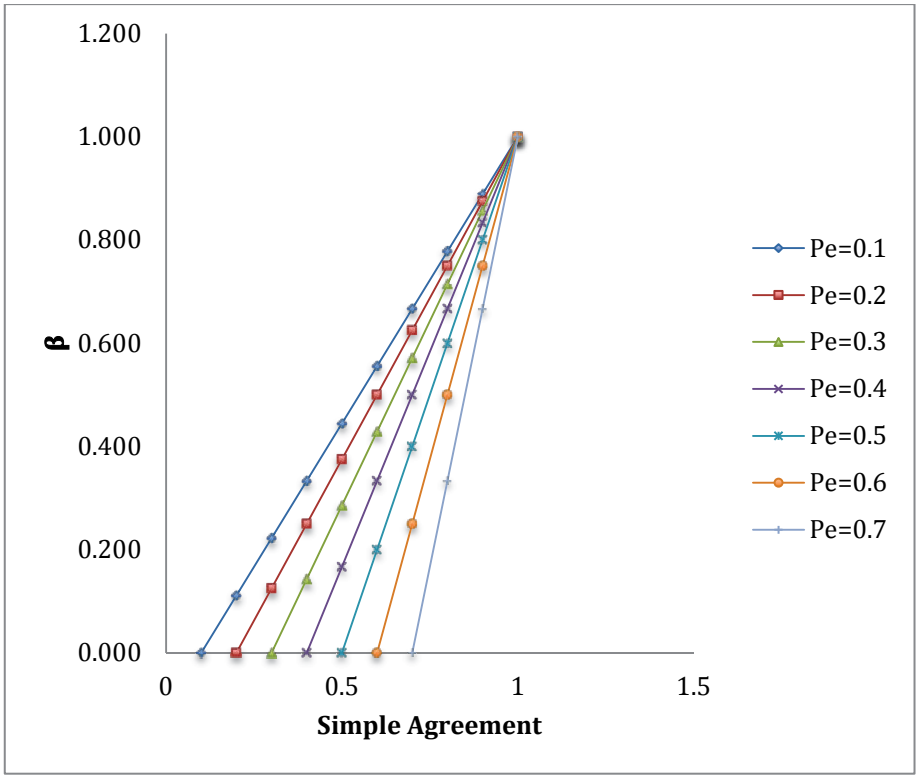

Figure 2.3.2.2: Graph of β statistic against simple agreement for various values of agreement by chance

In the same manner, as may be seen in Figure 2.3.2.3, the gradients of the lines of the difference between simple agreement with the agreement by chance removed increases as the agreement by chance increases. Thus small changes in agreement by chance can make a significant difference to determining agreement with chance removed.

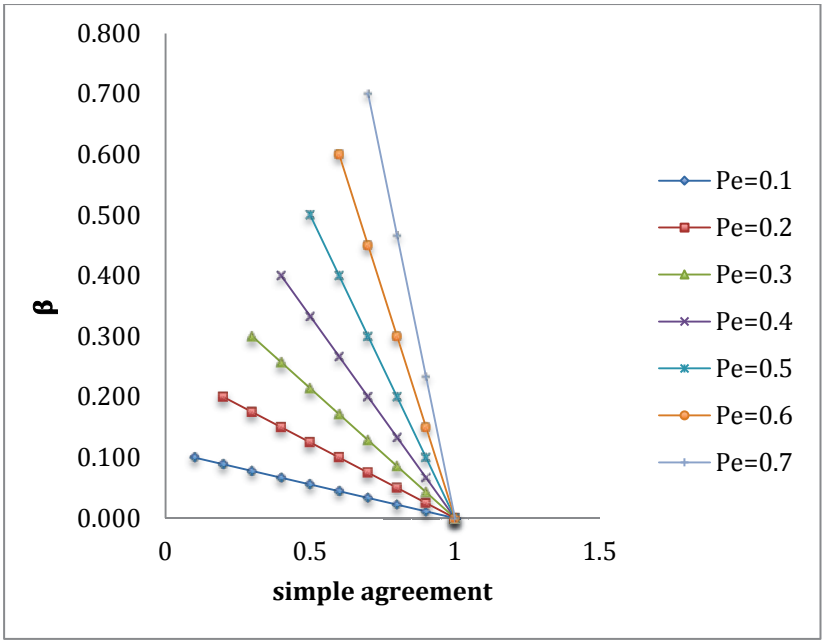

Figure 2.3.2.3: Graph of the difference between simple agreement and β as a function of simple agreement

Let α be the ratio between β and the simple agreement, namely,

$$\alpha = \frac{\beta}{\bar{P}}. \qquad (2.17)$$

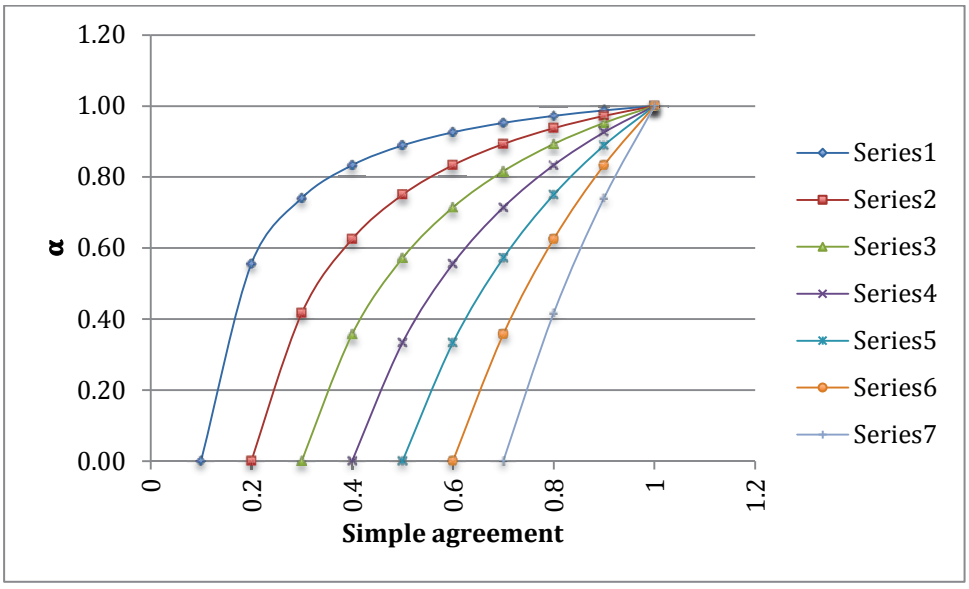

Figure 2.3.2.4: Graph of the ratio of agreement with chance removed to simple agreement against Simple agreement

It is readily seen in Figure 2.3.2.4 that as simple agreement decreases, the rate of decrease of the ratio of α to β grows. For example, when $\bar{P} = 0.7\ \bar{P}_e = 0.4,\quad \alpha = 0.71,$ when $\bar{P} = 0.6\ \bar{P}_e = 0.4\quad \beta = 0.56$ and when $\bar{P} = 0.5\ \bar{P}_e = 0.4,\ \alpha = 0.33,$ the difference between the values of α increases from 0.15 to 0.23. This, as mentioned above, means that at lower values of simple agreement the effect of removal of agreement by chance has a greater proportionate effect.

Upon reflection, it is obvious that an ordinal scale should be generally used to describe diagnoses, as diagnosis is typically never a simple 'yes' or 'no', or an 'either/or' situation and severity should be described. The papers cited in 2.3.1 and 2.3.2 restrict raters to answer within these constraints, an unrealistic limitation.

In the next section, papers that use ordinal scales to report the severity of symptoms or diagnoses and use weighted statistics to measure inter-rater agreement are evaluated.

2.3.3 Weighted Statistics

Only two investigations of diagnostic reliability in CM were found in the literature which utilised weighted statistics[26, 51]. In both cases Fleiss' Kappa was the underlying statistic. O'Brien *et al*[26] did not disclose the type of weighting used and Lo *et al*[51] used the quadratic approach. No papers were found that reported CM agreement using Gwet's AC2, the term used to describe the weighted AC1 statistic.

The inter-rater reliability results of both O'Brien *et al*[26] and Lo *et al*[51] for each diagnostic subgroup are presented in table 2.3.3.1.

Table 2.3.3.1 Weighted simple agreement and Kappa

Researcher	Topic	Sub-category	Choices available to raters	Number of practitioners	Simple Agreement	Simple Agreement pattern absent	Kappa
Lo[51]	Tongue diagnosis	Tongue color	5	12	NA		0.39 ±0.18
		Fur color	3	12	NA		0.83 ±0.04
		Fur thickness	3	12	NA		0.68 ±0.27
		Tongue Fissure	4	12	NA		0.72 ±0.18
		Red dot	4	12	NA		0.50 ±0.15
		Ecchymosis	4	12	NA		0.71 ±0.57
		Tooth marks	4	12	NA		0.59 ±0.24
		Saliva	4	12	NA		0.58 ±0.34
		Tongue Shape	3	12	NA		0.81 ±0.09
O'Brien[26]	Pulse diagnosis	Pulse depth	3	58	0.57		0.37
			5	58	0.43		0.35
		Pulse speed	3	57	0.61		0.4
			5	57	0.56		NA
		Pulse Strength	3	61	0.77		0.38
			5	61	0.62		0.37
	Hara diagnosis	Lung	2	62	0.58	0.06	0.08
		Spleen	2	62	0.48	0.11	0.07
		Liver	2	62	0.11	0.45	0.02
		Kidney	2	62	0.53	0.08	0.07
		Heart	2	62	0.00	0.95	NA

Lo *et al*[51] did not mention simple agreement, which was a stated criteria for inclusion into the literature review for the present work, but their results have been included as it was one of only two papers that utilised weighted statistics. Standard error values are given, which is a good addition to allow the correct interpretation of the statistical scores found by the author. Standard errors were sometimes quite large, of the nine diagnostic sub-groups reported, four had standard errors of 50% or greater associated with the Fleiss' Kappa value. As a consequence the values of Kappa presented in these cases have little value as an indication of agreement.

O'Brien *et al's*[26] study, the only paper found that met our criteria for review and reported Japanese acupuncture inter-rater agreement, did report simple agreement as well as weighted statistics. Interestingly, O'Brien *et al* when investigating hara diagnosis included simple agreement values when the practitioners agreed that a pattern was not perceived. Such an approach indicates that these researchers understood that agreement that a pattern or symptom was not present was a significant outcome. This was an important observation and was the only example found in the literature review that considered the non-selection of a symptom or pattern as an important aspect of agreement calculation. The fact that there are a small number of diagnostic options available in hara diagnosis enabled this viewpoint to be considered.

The weighted values of Fleiss' Kappa in all hara diagnoses were quite low, so that if these values were classified with the scale proposed by Landis and Koch[53], the agreement between raters would have been 'Poor'. None of the

Kappa values was greater than 0.1, in spite of the combined simple agreements of a pattern being present and non-present being all greater than 0.50. It seems that the combined agreements of pattern present and non-present were not used to determine the weighted Kappa value. This is demonstrated by the fact that the Liver hara pattern had a weighted Kappa value of 0.02, the lowest Kappa score recorded in O'Brien *et al's*[26] paper.

2.4 Validation exercise using the AC2 software

A hypothetical example of the use of the AC2 statistic will now be presented, which also will highlight its advantages over AC1, which does not take into account the proximity of scores between raters. Suppose that three raters are asked to score ten variables, A to J, on a six-point scale, 0-5. Their hypothetical evaluation assessments are presented in Table 2.4.1. Various scoring combinations have been developed by the present author so as to facilitate discussion about the factors affecting the calculated values of the agreement between rater with weighted and un-weighted statistics. For instance, the raters are in complete agreement that variable G was not present because they all allocated a value of zero. On the other hand there is only partial agreement concerning variable A, with raters 1 and 3 allocating a value of 1 to A the variable whilst Rater 2 allocated 4. Such differences in agreement will allow the change in outcomes to be evaluated when the same data is processed with weighted and un-weighted statistics.

Table 2.4.1 Scores for AC2 exercise

Variable	Rater 1	Rater 2	Rater 3
A	1	4	1
B	0	2	4
C	0	3	0
D	1	2	3
E	1	4	1
F	5	4	3
G	0	0	0
H	1	2	1
I	0	4	3
J	0	2	1

Table 2.4.2 Results with un-weighted simple agreement and Gwet's AC1 statistics

		Inference/Subjects		
METHOD	**Coefficient**	**StdErr**	**95% C.I.**	**p-Value**
Gwet's AC1	0.09	0.12	-0.19 to 0.367	4.895E-01
Simple Agreement	0.23	0.10	0.007 to 0.46	4.450E-02

As mentioned above, even a one-point difference in scores in the un-weighted approach is interpreted as a disagreement, while the differences in the scores in the weighted statistic approach causes agreement to undergo a reasonable reduction according to the proximity of the scores, as calculated with the following weighting table:

Table 2.4.3 Weighting matrix for linearly distributed weights calculate weighted agreement between two raters

	0	1	2	3	4	5
0	1	0.8	0.6	0.4	0.2	0
1	0.8	1	0.8	0.6	0.4	0.2
2	0.6	0.8	1	0.8	0.6	0.4
3	0.4	0.6	0.8	1	0.8	0.6
4	0.2	0.4	0.6	0.8	1	0.8
5	0	0.2	0.4	0.6	0.8	1

The linearly distributed weighting matrix used in the evaluation of the AC2 statistic is presented in Table 2.4.3. The horizontal row indicates the score assigned by one rater, while the vertical column shows the score allocated by the second rater. The value to be used as the weighting is the intersection of a row and column. If both raters allot the same value, then the rows and columns intersect along the diagonal of the matrix so that a weighting of 1 is assigned. If there is a difference of 1 that occurs between the two raters, then weighting of 0.8 is attributed during the evaluation of the AC2 value and so on.

Since the matrix in Table 2.4.3 is symmetric about the major diagonal, as mentioned above it does not matter which of the raters scores high and which scores low, the weighting is the same.

In order to illustrate the effect of weights, Table 2.4.4 has been developed. This table is Table 2.4.1 with two columns added. The average un-weighted and weighted agreements were calculated by hand for each variable and placed in the additional two columns

Table 2.4.5 Weighted and Un-weighted agreement in each variable

Variable	Rater 1	Rater 2	Rater 3	Average Un-weighted Agreement	Average Weighted Agreement
A	1	4	1	0.33	0.60
B	0	2	4	0	0.47
C	0	3	0	0.33	0.60
D	1	2	3	0	0.63
E	1	4	1	0.33	0.6
F	5	4	3	0	0.73
G	0	0	0	1	1
H	1	2	1	0.33	0.87
I	0	4	3	0	0.47
J	0	2	1	0	0.73
Averages				0.23	0.67

Table 2.4.4 clearly demonstrates the different outcomes that occur when weighted and un-weighted statistics are applied to raters' scores. For example, the simple unweighted agreement in variable E is 0.33, whereas the weighted simple agreement is 0.60. On the other hand, the simple un-weighted agreement in variable H is 0.33, whereas weighted simple agreement is 0.87. Such differences lead to different differences between the values of AC1 and AC2 statistics.

This is a profound difference which leads directly to the significant differences in calculating simple agreement as may be seen in the third column in Table 2.4.5. The un-weighted simple agreement obtained from the data in Table 2.4.2 is 0.23 whereas that obtained when using the weights in Table 2.4.4 is 0.68; essentially a threefold increase in agreement which would appear to more closely represent the data in Table 2.4.1.

The matrix in Table 2.4.3 was used to calculate the linearly weighted AC2 agreement shown in Table 2.4.5 based on the data in table 2.4.2. The values of simple agreement calculated by hand in Table 2.4.4 are in absolute agreement with the values given in Table 2.4.2 and in Table 2.4.5 for the un-weighted and weighted cases respectively, thus indicating that at least this calculation is performed correctly by the code used for the calculations.

It should be noted that the values of AC2 given in Table 2.4.5 as 0.23 is about 2.5 times the value of AC1 given in Table 2.4.2 as 0.09. Thus, whilst in this hypothetical case the AC1 statistic would indicate very poor agreement the AC2 statistic indicates that the agreement whilst not very good is much better. It can be argued therefore that the AC2 statistic is a significantly better measure of "true" agreement than the AC1 statistic and certainly quite superior to Fleiss' Kappa.

Table 2.4.5 Results with linearly weighted simple agreement and Gwet's AC2 statistics

METHOD	Coefficient	Inference/Subjects		
		StdErr	95% C.I.	p-Value
Gwet's AC2	0.23[1]	0.14	-0.087 to 0.556	1.327E-01
Simple Agreement	0.68	0.05	0.559 to 0.801	4.587E-07

1. The fact that Gwet's AC2 and the simple un-weighted agreement are the same is coincidental.

2.5 Strategies that lead to higher inter-rater agreement

Outlined in this section will be strategies that have been found by others to improve diagnostic agreement. This is an important section, as the goal of this thesis is to quantify and then improve inter-rater agreement in CM. Strategies that have been shown to be successful in improving diagnostic agreement need to be evaluated and considered for application to CM diagnosis. An obvious approach is to improve or increase rater training.

2.5.1 Rater Training

It has been consistently shown that rater training has had good effects upon the level of inter-rater reliability across diverse areas such as examination marking or interpreting radiology reports [60-64]. In CM the studies already reviewed by Sung *et al*[21] and Zang *et al*[22, 23] describe improvements in agreement after training. Mist *et al*[65] who used Fleiss' Kappa, also reported improvement diagnostic agreement after training in diagnosis of patterns. Mist *et al's* work was not included in the earlier literature review, as simple agreement was not presented in their paper. However, despite this improvement the complexities of CM still contribute to the difficulties.

2.5.2 Simplifying a diagnostic system

Investigations into what has caused diagnostic discord have taken place in psychiatry, a modality with similarly high numbers of diagnostic choices and low levels of inter-rater agreement as is the case in CM. Ward *et al*[66] found that the diagnostic structure in the psychiatric profession was associated with

63% of disagreements. The Diagnostic and Statistical Manual of Mental Disorders (DSM) is the recognised reference text for diagnosis in psychiatry, and has well publicised problems which result from the complexity of the manual[67-69]. The leading CM diagnostic definition reference is the World Health Organisation's International Standard of Terminologies[70, 71], also uses a DSM style expansive approach to defining disease states.

A diagnostic description in classic CM usually consists of a combination of terms, somewhat like a phrase. When people communicate, completely different combinations of words can be used to describe nearly the same thing. Similarly, substitutions or combinations of CM patterns can be used to describe essentially the same condition of a patient. For example Liver Qi Stasis, a very common CM diagnosis is also described as Wood (Liver) invading Earth (Spleen), general Qi Stagnation, Qi Stagnation in the Gall Bladder channel, or other patterns.

As mentioned above, herein lies a major problem in expressing a diagnosis in the classic CM method; exact phonetic agreement is unlikely as there are many possible choices of diagnosis, which express the same or closely related disease states. This was intimated in Zhang *et al's* studies[22, 23] earlier summarised in 2.3.1, where a loosening of the diagnostic agreement criteria was shown to lead to improvements in agreement. The strategy that seems to have been generally adopted in the CM research community to reduce the effect of this difficulty has been to investigate a single disease state[22-24, 27, 28, 31, 57, 72-76], or develop and validate questionnaires for a single diagnostic

facet[59, 77-80]. This approach has the tendency to inflate agreement as the number of options available is drastically reduced.

2.5.3 Questionaires

There have been attempts to develop questionnaires to improve diagnostic reliability in CM and in other modalities. This section briefly summarises the questionnaires that have been published, especially in relation to CM.

Many questionnaires have been developed for modern biomedicine and diseases. For example, there are more than eighty questionnaires used to describe back pain with five most frequently used[81, 82]. Questionnaires such as the SF36[83] is a self-reported wellness measure that is widely used in medical research and has been validated[84, 85]. The SF36 has been reported to be useful for measuring changes in a subject's health condition, but users of this questionnaire are cautioned that it may not be accurate as a definitive, objective measure of the health of a subject[86].

Adopting a similar theme are diagnostic questionnaires in CM that are generally limited to one condition such as Blood Stasis[79] , or two opposing diagnostic facets such as Heat and Cold[80]. It seems that a questionnaire addressing the whole CM diagnostic framework has rarely been attempted.

Park *et al*[79] published a questionnaire for diagnosis of blood stasis by identifying 48 variables and subsequently conducting Delphi panels with 17 practitioners of more than five years experience, to reduce the number of

variables to 20, each to be scored on a Likert scale of 0-6. Internal reliability of 0.790 was reported with Chronach's α, which is above 0.700, the score considered as acceptable[87]. Construct validity was examined using principal component analysis with varimax rotation[88].

To determine validity of the questionnaire as compared to the diagnoses of practitioners, a group of 61 patients, 25 male and 36 female, having completed the questionnaire were then diagnosed independently by three clinicians with more than five years' experience. The clinician's agreement was only considered valid if all three practitioners agreed, and the level of agreement between the three clinicians in whole groups was not reported. The total agreement between the clinicians was compared with the results obtained with the questionnaire using Randolph's free-marginal Kappa calculator. The result obtained was a Randolph free-marginal Kappa of 0.825 which is deemed 'Almost Perfect' agreement according to Landis and Kochs' scale of chance-removed agreement interpretation[53], with no standard error disclosed.

Randolph's free-marginal Kappa has had minimal uptake and the paper[49] that describes its theoretical foundations was self-published and not peer reviewed. While one paper was found in Korean[89], no one has published an evaluation of the essential differences, if any, between Gwet's AC1 and Randolph's Free Marginal Kappa in English as yet, so the differences between these two statistics are unclear. This paper would also have benefited from reporting the level of agreement between the practitioners and

the standard error of the inter-rater agreement between the practitioners. Despite the shortcomings in the reporting detail and choice of inter-rater statistic used to describe agreement with these data, the experimental methodology is interesting and the experiments seemed to be well performed

Interestingly, Park *et al* also elucidated a relationship between blood stasis and heart rate variability, which provides an objective measure for the diagnosis of blood stasis.

Park, with slightly different teams also produced a similarly derived twenty-seven item questionnaire for yin deficiency[77] and a twenty-five item questionnaire for phlegm patterns[78], but did not present validation against the opinions of practitioners in either paper. The yin deficiency paper did not mention agreement between practitioners at all and after disclosing the inclusion of three practitioners being used in the phlegm study to determine pattern agreement, there seemed be no report of the inter rater reliability between the questionnaire and the practitioners. Both of these papers would have benefited from the inclusion of details of each questionnaire's reliability when compared with practitioners.

Ryu *et al*[80] reports a twenty-item 'yes or no' questionnaire for Cold-Heat patterns, with ten items for each pattern and published the questions as an appendix to the paper. Chronbach's α was used to determine internal consistency. Chronbach's α was 0.579 for the Cold questions and 0.718 for the Heat questions. It has been suggested[87] that values of α>0.5 are

acceptable, although scores ideally should be α>0.7. Two groups were trialed with the questionnaire. Discriminate validity was assessed by independent sample Student's *t* test, against diagnosing doctors. Greater than 90% classification accuracy was obtained in both groups. Ryu *et al's* paper would have benefited from the inclusion of chance-removed statistics and a weighted scoring system for the questionnaire.

Schyner *et al*[59] in 2005 developed a structured interview called the TEAMSI TCM, which is meant to be prescriptive instead of descriptive. In other words, the questionnaire is populated and the result obtained is meant to be a normative influence on the diagnosis given by the interviewer. The TEAMSI TCM is meant to increase inter-rater consistency, not replace practitioners' diagnoses.

The questions included were refined through a Delphi Panel[90] of ten expert practitioners and a validation against practitioners was announced as being underway in the introductory paper[59]. This validation seems to not be publicly available, nor appears to be the questionnaire, which suggests that the TEAMSI TCM questionnaire is probably a commercial product that can be used only under licence and not available in the public domain. TEAMSI TCM was recently included in a study by Grant *et al*[57] where TEAMSI was endorsed as a good addition to their study stated as validated, but no reference was given to support this assertion.

2.6 The Diagnostic System of Oriental Medicine

One questionnaire was found that did attempt to investigate the disease state of a patient using the entire system of CM diagnostic patterns. The instrument was titled the “Diagnostic System of Oriental Medicine” (DSOM). The DSOM is a Korean diagnostic questionnaire developed by Lee[91], designed initially to diagnose and investigate women’s reproductive health from a Korean CM perspective. The DSOM was used by several diagnostic investigations around 2007[92-95] and did not seem to be used again until 2014[96], where the Descriptors used to define CM diagnoses within the DSOM were disclosed. The format of presenting the results obtained with the DSOM questionnaire, hereby designated as DSOMf, given in that paper was found to have the potential to resolve the semantic issues of describing a diagnosis defined in section 2.1.

Surprisingly, the DSOMf consists of scores ascribed to only 16 diagnostic Descriptors, namely:

- Five organ/elements; heart, spleen, lung, kidney and liver,
- Five pathogenic factors; heat, cold, damp, phlegm and dryness,
- Four deficiencies; qi, blood, yin and yang, and
- Two stagnations; qi and blood.

The Descriptors in the DSOMf are akin to Lego pieces that are combined to describe a patient’s condition. Either multiple or single pattern CM diagnoses can be portrayed within the standard format of the DSOMf, an advantage over the current system used by CM.

Three different measures were made available by Lee that reported severity of pathology with the DSOMf within each Descriptor after the questionnaire data was processed. These were Raw Score, a number between 0 and 100, a Weighted Score, a number between 0-10 and Weighted Rank, one of four possible combinations of LL, LH, HL or HH. These were used by the DSOM to report the certainty of the result and only made available if data was submitted to Lee and have not been used in the processes developed within this thesis.

Unlike the contemporary diagnostic format used by most practitioners, in which, as has already been mentioned a number of times, over one hundred patterns are available for selection, each of the 16 descriptors are equally relevant whether scored or not and contribute to evaluating the overall diagnostic picture of the patient. The DSOMf also lent itself to the easy application of chance-removed inter-rater agreement statistics discussed earlier.

Since there are no near duplications and the non-selection of a descriptor means that the pathology is not present, so that the DSOMf has the potential to allow true comparisons to be determined between practitioners making diagnostic choices. This approach of recording systematically the material for arriving at a diagnosis warrants further investigation.

However, there are some difficulties associated with the use of the DSOM questionnaire. The written questions had to be translated from Korean. This

meant that whereas speakers of that language easily understand some expressions in Korean, the exact meaning was difficult to establish in English. The translated questions therefore sometimes seemed “challenging”; for example it is not clear what is implied by “I feel as if my brain is shaking or vibrating”. The full version of the translated questionnaire is presented in Appendix 1. There were 151 questions, distributed into the following 17 categories in Table 2.6.1.1:

Table 2.6.1.1 Question categories and number of questions in the DSOM questionnaire

facial symptoms	7
appetite	4
thirst	6
digestion	12
condition of the stool	3
a tendency to diarrhoea	10
perspiration	6
sensitivity to heat and cold	15
my emotional state	12
body pain	24
dizziness	4
fatigue	10
sleep	4
skin, hair and nails	7
limbs	3
other questions	19
urination	5
Total	151

Questions were not identified as to which of the sixteen diagnostic descriptors they contributed to, nor how much weight was assigned to each question. Surprisingly there are no questions relating to a woman’s reproductive system despite the fact that the questionnaire had been originally devised to investigate woman’s health in the CM context. The state of a woman’s menstrual cycle, how regular it is, or the quality of the menstruation can be

used to diagnose a patient's health. It seems unusual that this direct line of questioning was not adopted.

Question numbering does not inspire confidence and is seemingly a work in progress, with an apparently random pattern in the numbering sequence. Twenty-five questions were grouped in 'a' and 'b' divisions of the same number and many question numbers were missing. The highest question number in the questionnaire is 169. The peculiar number sequence present in the questionnaire seemed to point to a process of some questions being removed and/or added in later editions after a degree of testing.

It seems that there was an attempt to make the questionnaire broader and include men's health, as evidenced by the fact that differing numbers of questions (152 for women and 149 for male) are mentioned in a paper dated 2011[91] by Lee *et al.* This also points to an evolution of sorts taking place in the deciding on the questions included in the DSOM questionnaire.

Some questions were as would be expected as part of a CM diagnostic questionnaire, for example "On rising in the morning I have diarrhoea", while others seem a bit odd, such as "I feel as if my brain is shaking or vibrating" mentioned above. The questions were not referenced to any textbook, so their selection for inclusion seemed to be at the discretion of Lee and her co-researchers.

Each question was answered with one of five possible options, as outlined below:

① Not Applicable ② Weak ③ Moderate ④ Strong ⑤ Very strong

or sometimes

① Never ② Hardly ever ③ Sometimes ④ Frequently ⑤ All the time

Some questions would be expected to have higher weightings than others, and therefore greater or lesser diagnostic significance, but the weighting algorithm is not known. Further, even if the weighting matrix were known, the method used to derive the diagnoses from the questionnaire to produce the results presented with the DSOMf has not been identified. As a consequence the results needed to be processed by Lee in order to obtain diagnostic scores from the raw questionnaire data. Finally, as far as currently known, the process has not been validated in the English-speaking world. All these factors provide serious impediments that hinder the DSOM questionnaire from being an instrument that could be generally used to determine CM diagnoses, particularly in the English speaking world. However, the DSOMf for recording as well as comparing CM diagnoses from different practitioners is an interesting and important contribution.

2.7 Summary of Literature review

As far as the present author is aware, there have been no studies of open populations with unrestricted health statuses in Chinese Medicine. Studies were focussed exclusively on a specific Western medical disease, with strictly limited CM diagnostic options or a plethora of 'yes or no' options within diagnostic sub-groups were made available to the practitioners.

The qualifications of the practitioners used to diagnose were almost always described and when disclosed were invariably of a high standard, but somewhat in contradiction the variability in diagnosis was often attributed to differences in the level of experience of raters. Blinding of the raters' diagnoses was commonly reported. Training and the use of guiding questionnaires were sometimes used to improve agreement between the raters. The researchers' comments as to the levels of agreement were always appropriate.

Incorrect Kappa statistics were utilised in most studies. Interestingly, none of the authors gave any reason or even commented on the fact that there were often such large differences between simple agreement and agreement after chance agreement was removed. Surely, this observation supports the view that a confounding factor had to be present, probably differing marginalities of data from various research projects. It seems therefore necessary to use a measure such as Gwet's AC1, which just like Fleiss' Kappa for Fixed Marginal Data can also be used with Free Marginal Data; the most likely type of data encountered in research.

The approach used to investigate agreement in the studies reviewed, were often many small micro tasks, such as an attribute of the tongue or face color, which are only part of the diagnostic process. This approach misses a vital point in that these microanalyses have to be assembled in some way to determine a diagnosis. The emphasis that a practitioner places on each diagnostic facet will almost certainly differ in detail from others although they may arrive at the same diagnosis.

Even more of concern, is the fact that there has not been any discussion of the apparent ‘leap of faith’ that moving from many sub-diagnostic units to full diagnosis entails the assumption that the same weighting of each symptom is given by all practitioners. The breaking up of a complex task into a number of smaller units and investigating each one individually is a positive and useful process, but the integration of these diagnostic units into a diagnostic outcome is yet another matter and may be a difficult but essential aspect of the diagnostic process to investigate. This is another research area that is a study in itself.

Certainly, the simplified and fragmented approach used to evaluate agreement, does not reflect the situation that confronts practitioners every day in contemporary CM practices. In general new patients arrive at a clinic without necessarily having indicated the reason for requesting an appointment and certainly without having been pre-screened or diagnosed by the diagnosing practitioner. It seems that a ‘laissez faire’ approach is adopted

since each practitioner views each case as unique, assigns the number of diagnostic labels as he/she considers necessary without any limit within whatever general diagnostic theory that the practitioner deems appropriate.

The inherent assumption of a ‘correct diagnosis’ being the one that is capable of being repeated by many practitioners is also a point of conjecture. Unless there is a method of determining the “correct” diagnosis, such as a conclusive physical measurement that can be used, there is no way of determining whether the consensus diagnosis, that is the mean of many raters, which by default becomes the diagnosis is indeed the correct diagnosis. In the absence of such conclusive physical measurements, an acceptably high consensus of diagnosing practitioners is the appropriate basis to commence investigations.

‘Big data’ is expanding rapidly in all areas including medicine and para-medicine[97]. Good quality clinical records will be required to take advantage of this exciting opportunity to effectively mine this information. One of the many opportunities that big data offers is the potential to compare treatments and their effectiveness to diagnoses. These investigations will rely on non-fragmented data. Evidence of CM diagnostic consistency will be essential. A standardised diagnostic format adopted widely that allows diagnoses and treatments of many practitioners to be pooled will be critical for the accumulation of high quality data.

O’Brien and Birch[24] sum up matters appropriately;

> ‘Reliability of pattern diagnosis has been found to be variable across types of practice and a range of diseases’.

The same review article fittingly concludes that:

> ‘Until diagnostic data collection and pattern diagnosis are shown to be reliable, there can be little justification for inclusion in clinical studies of TEAM[3]. Therefore, it is important that researchers develop strategies to improve reliability’

These comments and the results in the cited studies both confirm that something must be done to improve diagnostic agreement in CM as a priority, certainly before contemplating the use of data mining.

In conclusion, there is a lack of any investigations into an ‘open population’. The wrong statistics were almost always used to determine agreement, together with a frequent failure to report standard error or confidence interval statistics. Mostly two or sometimes three raters, the minimum necessary to allow a comparison of diagnoses was used for what seems to be research convenience. The DSOMf was identified as a possible means to facilitate superior agreement. Rater training was also found to be associated with improvement in diagnostic agreement.

[3] Traditional East Asian Medicine

Accordingly, an investigation into diagnostic reliability that might be obtained from an open population of subjects, with higher numbers of practitioners who recorded their diagnosis with the DSOMf utilising the appropriate statistic, the AC2 should be initiated. This investigation would seem a logical step towards attempting to benchmark agreement with appropriate statistics, with what seems to be a promising diagnostic format.

Chapter 3 Investigation of DSOM Diagnostic Reliability

An investigation of diagnostic reliability using the DSOMf was carried out. Broadly, the inter-rater diagnostic agreement between five practitioners who would examine a reasonable sample of subjects from an open population was to be determined with each practitioner using the DSOMf.

An open population in this instance is defined as a group of subjects who were not restricted by age or gender or disease. This group of subjects was meant to represent the sort of patient that might walk into any CM clinic in the Western world, specifically Australia. This experiment is therefore an attempt to quantify the levels of agreement that might take place in this country's contemporary CM clinical setting.

The results obtained from the five practitioners were also compared with the diagnosis derived from the DSOM questionnaire provided when the questionnaire data was sent to Lee and processed.

The procedures and all the details regarding this data collection are presented in this section: from data collection, to details of subject and practitioner recruitment.

3.1 Data Collection Details

Each subject filled in the DSOM questionnaire and saw five practitioners in turn for interviews. Appendix 1 contains the complete DSOM questionnaire.

The sequence of the data collection was as follows. Upon arrival the subject would be greeted and would fill in a subject consent form and receive a subject information form. Full details of these forms are shown in Appendices 2 and 3. Then each subject was given a copy of the DSOM Questionnaire to fill in. After completing these forms the subject were ready to see the practitioner for interviews.

The practitioners, after interviewing each subject, filled in a diagnostic data sheet that had on the top the subject's number and the diagnosing practitioner's number. This allowed each practitioner's diagnosis to be compiled and accounted for. The practitioners rated each of sixteen Descriptors with a Likert scale[98]. Each Descriptor was rated between 0-5, where zero indicated the pattern was absolutely not present and five indicated it was there maximally. The DSOMf diagnostic data sheet used in this experiment is presented in Appendix 4.

Quasi-randomisation of subject assessments by the practitioners occurred in the following manner; when a subject completed an interview with a practitioner, they saw the next available practitioner. This procedure also increased workflow efficiency.

The practitioners were asked to diagnose in the same way that they would in their normal clinical setting. The practitioners and subjects were instructed not to discuss their interviews during the course of the data collection.

3.3.1 Setting

Data was collected at the outpatient Chinese medicine clinic at UTS, Harris St Ultimo, on a weekend when the clinic was closed to the public.

3.3.2 Subjects

A total of 43 subjects attended; of an average age 44.3 years with 23 females and 20 males. One subject's data was discarded due to a missing practitioner's DSOMf diagnostic sheet, leading to 42 subjects data being analysed. The subjects were either patients at the primary investigator's clinic, all with reasonable levels of health, or healthy, competitive masters cyclists known to the primary investigator. Since all the subjects were known and, in many cases, were patients who had been treated by the primary investigator for extended periods, the present author did not perform any of the diagnosis in this experiment. All subjects had been cleared of serious health issues by their general practitioner. Each subject received a reimbursement of $20 for participation in the study, which involved completing questionnaires and being interviewed by five CM practitioners. No treatment was offered to the participants.

Data was collected in the following manner: days one and two were consecutive days, whereas day three occurred six months later. After processing the data from days one and two, it was found that a significant proportion of the subjects were virtually symptom free, as reflected by their high levels of fitness from competitive cycling. As a consequence, the subjects seen on day three were intentionally selected with the characteristics

of lower levels of health than those seen on days one and two, to increase the representation of unwell subjects in the collected data. The diagnosing practitioners were not made aware of the change in choice of subjects in day three.

3.3.3 Practitioners

Each practitioner utilised as raters in this study had at least five years' full-time experience and trained either at UTS or in China. All practitioners received an honorarium of $100 per day for participation in the study. A total of nine practitioners of Traditional Chinese Medicine were used in this study, in various combinations on each of three days' of data collection. The use of five practitioners provided the opportunity for greater robustness in exploring inter-rater reliability than had been reported previously. This was identified as a weakness chapter in all studies examined in the previous. Thus, this study breaks new ground in the investigation of inter-rater reliability simple by using a larger number of raters.

3.3.4 Data Processing to Determine Agreement

The collected data was entered into Excel™ for data processing. Gwet's Excel™ based program[52] was used to determine linearly weighted simple and AC2 chance-removed agreement and the associated standard errors in a similar format as previously described in section 2.4. An example of how the data was processed is given below.

The data was organised with the scores of 0-5 each practitioners gave to each Descriptor forming rows, in the format given in table 3.3.4.1. In smaller data sets, subgroup analysis was used and a test group containing the full score range was included to ensure that the correct weighting table was always employed.

Table 3.3.4.1 Data processing format for the calculation of agreements using the DSOMf

Subject	Descriptor	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
1	Cold	1	0	1	0	3
1	Heat	0	0	0	0	2
1	Damp	3	4	0	4	2
1	Dry	0	0	0	0	0
1	Phlegm	0	0	2	4	4
1	Qi Xu	3	3	4	3	1
1	Blood Xu	0	3	0	3	4
1	Qi Stag	1	3	2	0	3
1	Blood Stag	0	0	4	0	0
1	Yin Xu	0	0	0	4	0
1	Yang Xu	1	0	1	0	4
1	Liver	0	0	1	0	0
1	Heart	0	0	0	0	0
1	Spleen	4	3	0	4	4
1	Kidney	0	0	5	4	4
1	Lung	0	0	0	0	0
2	Cold	0	0	0	0	0
2	Heat	0	3	3	0	4
2	Damp	0	0	0	0	0
2	Dry	2	0	2	0	4

Table 3.3.4.1 Data processing format for the calculation of agreements using the DSOMf (continued).

Subject	Descriptor	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
2	Phlegm	0	0	0	0	1
2	Qi Xu	0	0	0	0	1
2	Blood Xu	4	0	2	0	3
2	Qi Stag	4	4	4	3	3
2	Blood Stag	0	4	3	3	4
2	Yin Xu	3	0	4	0	4
2	Yang Xu	0	0	0	0	0
2	Liver	4	4	5	4	4
2	Heart	0	0	0	4	1
2	Spleen	2	0	0	0	3
2	Kidney	1	0	0	0	1
2	Lung	0	0	0	0	0
3	Cold	0	0	3	0	0
3	Heat	0	0	0	2	4
3	Damp	3	0	2	4	0
3	Dry	0	0	4	0	1
3	Phlegm	0	0	0	1	0
3	Qi Xu	3	2	2	3	3
3	Blood Xu	1	0	0	0	3
3	Qi Stag	1	3	2	0	0
3	Blood Stag	0	0	0	0	3
3	Yin Xu	0	0	0	0	3
3	Yang Xu	0	0	0	0	0
3	Liver	2	3	3	0	0
3	Heart	2	2	2	1	4
3	Spleen	4	2	3	3	2
3	Kidney	0	0	0	0	3
3	Lung	0	0	0	0	0

Table 3.3.4.1 contains the data collected from the first three subjects on day one of the data collection. Gwet's program was used to calculate the agreement that occurred in each row and the inclusion of rows into the data table enabled the agreement of all data included to be calculated in a combined manner. In this way for instance, all the raters' scores allocated to all the Descriptors of one or a group of subjects.

Altering the inclusion criteria to the data table enabled great flexibility in examination of inter-rater agreement of these data. Various criteria, such as each day, each Descriptor, total scores allocated to individual subjects or Descriptors from all interviewing practitioners, were used as sorting criteria to explore many characteristics of the data. The inter-rater agreement results derived from the diverse criteria are reported in chapter 4.

Chapter 4 The DSOM Data Collection

The DSOM data set obtained using the DSOMf by five practitioners examining 42 subjects, are summarised and discussed in this chapter. Statistical analyses are performed on the data to determine the levels of agreement and the results discussed in some detail.

4.1 Data Recorded by the Practitioners

The overall score choices of the raters and then the descriptor selections will first be summarised.

The number of times each score was selected by practitioners on all days are Presented in table 4.1.1.

Table 4.1.1 Score selections of the practitioners with the DSOM data set.

Score	0	1	2	3	4	5	total
Selected	1990	256	286	409	312	107	3360
Percent Selected	59%	8%	9%	12%	9%	3%	100%

By far the most often chosen value by the practitioners was 0, which had been select 59% of the time as the state of health for the various Descriptors for the subjects. The second most frequently selected value was 3, with the scores 1,2 and 4 each was used on about 10% of occasions whereas the quantity 3 was used marginally more often at 12%. The highest score available to the practitioners was 5 and was intended to indicate that the syndrome represented by a Descriptor was present at the highest possible level, was

employed on only 3% of the total number of selections made by the practitioners. The slightly more frequent use of 3 than 1, 2 or 4 scores selections suggests that raters may be more likely to score a descriptor 3 as a standard selection if they thought there was a problem with the patient within this Descriptor. The low frequency of 5 scores perhaps implies either a reluctance to score maximally, which could be seen as a common conservative trait, or that very few of the subjects presented with extreme health problems.

The number of occasions each descriptor was allocated a score and the value of that score is presented in Table 4.1.2. The descriptors are presented in order of the frequency of their selection, with the most often used descriptor presented first and the descriptors with least number of selections placed last.

Table 4.1.2 Descriptor Score selections

Descriptor	0	1	2	3	4	5	Selected
Qi Stag	33%	10%	15%	21%	16%	5%	67%
Spleen	33%	10%	16%	20%	15%	5%	67%
Qi Xu	34%	10%	16%	19%	16%	5%	66%
Liver	35%	8%	10%	20%	17%	10%	65%
Kidney	49%	6%	8%	11%	17%	9%	51%
Heat	52%	9%	10%	16%	11%	3%	48%
Yin Xu	62%	7%	6%	12%	12%	1%	38%
Damp	63%	9%	7%	13%	7%	1%	37%
Blood Stag	65%	6%	10%	12%	6%	0%	35%

Table 4.1.2 Descriptor Score selections (continued)

Heart	67%	8%	5%	10%	6%	4%	33%
Lung	70%	8%	4%	10%	6%	2%	30%
Blood Xu	71%	6%	8%	8%	6%	1%	29%
Yang Xu	74%	7%	6%	7%	3%	2%	26%
Cold	77%	5%	7%	8%	2%	1%	23%
Dry	80%	8%	3%	4%	4%	0%	20%
Phlegm	81%	6%	5%	3%	5%	0%	19%
All Descriptors	59%	8%	9%	12%	9%	3%	100%

The well-known Descriptors; Qi Stagnation, Spleen, Qi Xu and Liver are all shown in Table 4.1.2 to have been used on about 65% of occasions, which could indicate that these Descriptors are most likely to occur in an open population. Kidney was employed on significantly fewer instances namely, 51%, to make up a top five Descriptor cohort. At the other end of the spectrum, Dry and Phlegm were the least often selected descriptors being chosen on only about 20% of occasions suggesting that they are most likely not to be found in an open population.

4.2 Simple and AC2 Agreement

Linearly weighted simple and AC2 agreement between the raters diagnosing all 42 subjects and recording their diagnosis using the DSOM is reported below in table 4.2.2.

Table 4.2.1 DSOM agreement using all 16 descriptors.

Simple Agreement	AC2
78 ±0.01	0.60 ±0.02

AC2 linearly weighted agreement attained the threshold of Substantial agreement according to the Landis and Koch[53] scale across all Descriptors. The standard error is quite low, suggesting the result is quite robust. When these results are compared with those that were found in the literature, this much improved agreement should be recognised as a positive outcome. A number of factors need to be recognized as possible causes for the improved inter-rater agreement.

As mentioned above in section 2.3.1.2, agreement usually declines when higher numbers of raters were used. As previously stated, the DSOM study used five raters, a larger number than any studies found in the literature, but overall inter-rater agreement between the raters who used the DSOM, rather than being lower than that found in earlier studies, was higher than had been previously achieved. Indeed, the agreeemnt with agreement by chance removed, called AC2 in Table 4.2.1, is very high.

The present results could only be compared with only two data sets[27, 28] found in the literature which used the AC1 statistic in CM diagnostic studies. These two studies were limited to stroke patients with only two diagnostic option available to two raters only in many the diagnostic sub-units investigated. In each study, the average AC1 agreements of all the diagnostic

sub-units were a Substantial 0.71 and 0.74 respectively. This very high inter-rater agreement could be simply due to the small number of raters and certainly to the small number of choices given to them. This certainly was not the case in the present investigation.

The DSOM data set was obtained from an open population which would be likely to produce a downwards pressure on agreement. A higher number of diagnostic options available to the practitioners usually means, that agreement is the less likely. Once again this did not happen in the present research.

On the other hand, the fact that agreement was also calculated in the DSOM data set with all scores included. Since, as shown in Tables 4.1.1 and 4.1.2 many raters scored many Descriptors as zero, the calculated agreement would have had an enhancing effect upon the agreeement calculated. This possibility of agreement upon non presence of patterns was reported in O'Brien *et al's* Toyohari paper[26] in simple agreement form but not incorporated into the chance-removed agreement calculations. The use of weighted statistics would also have provided improvement in agreement as well.

The DSOM data set inter-rater agreement just presented is the only example known of CM agreement being investigated in an open population with a larger number of practitioners. As such, it is an important step towards

benchmarking rater agreement between multiple raters within open populations in the CM profession.

4.3 Total Patient Pathogenic Score and Agreement

The total of the scores ascribed to each descriptor in each subjects' diagnoses were used to calculate a Total Pathogenic Score (TPS). Since different levels of agreement could be expected to depend on the complexity of subjects' health status, three wellness groups were formed on the basis of their TPS; each group containing 14 subjects, or one third of the total number.

The TPS can be used as a generalised wellness measure of a patient within CM terms. The TPS can also be used to reliably track *changes* in the overall health of a patient; a lowering in the TPS indicates an improvement in health and a rise in the TPS a decrease in health. Score changes in a subject's health status recorded by the same practitioner should be more useful for the determination of health changes than absolute descriptor scores[84], due to possible practitioner scoring bias.

The average TPS in each of the individual descriptors is presented in the first column of Table 4.3.1. As a comparison, the percentage of occasions each descriptor was scored 1 or greater was previously presented above in Table 4.1.2, is reproduced in the second column of table 4.3.1.

Table 4.3.1 Average Total Practitioner Score and percentage scored one or above for each Descriptor.

Descriptor	Av TPS	Selected
Spleen	1.41	67%
Qi Xu	1.37	66%
Qi Stag	1.10	67%
Kidney	0.99	51%
Liver	0.89	65%
Yin Xu	0.69	38%
Heat	0.67	48%
Blood Stag	0.57	35%
Damp	0.46	37%
Yang Xu	0.36	26%
Lung	0.33	30%
Blood Xu	0.31	29%
Cold	0.21	23%
Heart	0.21	33%
Dry	0.20	20%
Phlegm	0.11	19%
All Descriptors	0.62	41%

When compared with the Descriptor rate of score value selection mentioned in section 4.1.2, the ‘top five’ Descriptor selections are the same as selected by two methods. The only difference is the order, with Kidney being scored higher than Liver in average TPS value. Phlegm clearly had the lowest

average TPS of all, while Dryness was next lowest. The 0.94 correlation coefficient between these two approaches to characterising the weight of descriptor selection was very high.

It can therefore be concluded that using two methods of determining Descriptor selection that Spleen, Qi Xu, Qi Stagnation, Liver and Kidney were the most heavily utilised, while Dry and Phlegm were the least utilised.

4.4 Practitioner Agreement in the Three Wellness Groups

Practitioner agreement within Wellness groups using linear weighting is presented in table 4.4.1.

Table 4.4.1 Agreements in each Wellness group.

Groups	Simple Agreement	AC2	TPS Average
Most Well	0.85 ±0.01	0.77 ±0.02	49
Intermediate	0.76 ±0.01	0.57 ±0.03	93
Least Well	0.73 ±0.01	0.42 ±0.03	132
All Groups	0.78 ±0.01	0.60 ±0.02	91

Whilst the Most Well cohort falls into the Substantial agreement interpretation on Landis and Koch's scale[53], the Intermediate and Least Well cohorts would be rated Moderate. The Substantial agreement between raters of the Most Well group, arguably the lowest acceptable level, is a positive outcome. It indicates that agreement when patients are well is robust using the DSOM.

This is a good outcome and supports the view that the CM practitioners are able to correctly identify high levels of wellness utilising the DSOM.

The comparatively reduced Moderate agreement found in the Intermediate and Least Well groups warrants further investigation.

4.5 Fleiss' Kappa revisited

A minor digression will now take place, to further verify the assertions made in the Literature Review Chapter regarding the inappropriateness of using Fleiss' Kappa statistics with free marginal data. The linearly weighted Fleiss' Kappa in each of the wellness groups is now presented in Table 4.4.1, alongside the previously reported simple agreement and AC2 results. As the marginality of these data is known, the differences in score distribution should have made an impact in the Fleiss' Kappa result when compared to that derived with the AC2 and simple agreement statistics.

The score distribution in each of the three Wellness Groups is presented below in Table 4.5.1 to demonstrate the marginality of each wellness group.

Table 4.5.1 Score selections by the practitioners in the three wellness groups

Score	0	1	2	3	4	5
Most Well	73%	9%	7%	7%	3%	1%
Intermediate	59%	6%	9%	14%	10%	3%
Least Well	46%	8%	9%	16%	15%	6%

The proportions of score selection by the raters change in the way that might have been expected. Zero reduces from 73% in the Most Well Group to 46%

in the Least Well, whereas scores of three and above increase from the Most Well to the Least Well.

While none of the groups' data is fixed marginal, i.e. uniformly distributed, the Most Well group has the greatest inequality in the distribution. This should and, does, lead to differences between the AC2 and Fleiss' statistics, as may be seen in Table 4.5.2. and Figure 4.5.1. The noticeable differences in the values of the AC2 statistic and Fleiss' Kappa in table 4.5.2 are clearly delineated in the last right hand column in which the discrepancy between the AC2 scores and Fleiss' Kappa are presented. The standard error of the difference between the AC2 and Fleiss' Kappa scores was calculated as the root of the sum of the squares of the errors.

Table 4.5.2 Linearly weighted Fleiss' Kappa and AC2 agreement and Differences in the three Wellness Groups.

Wellness Groups	Simple Agreement	AC2	Fleiss'	Difference between AC2 and Fleiss'
Most Well	0.85 ±0.01	0.77 ±0.02	0.22 ±0.03	0.55 ±0.04
Intermediate	0.76 ±0.01	0.57 ±0.03	0.25 ±0.03	0.27 ±0.04
Least Well	0.73 ±0.01	0.42 ±0.03	0.29 ±0.03	0.13 ±0.04

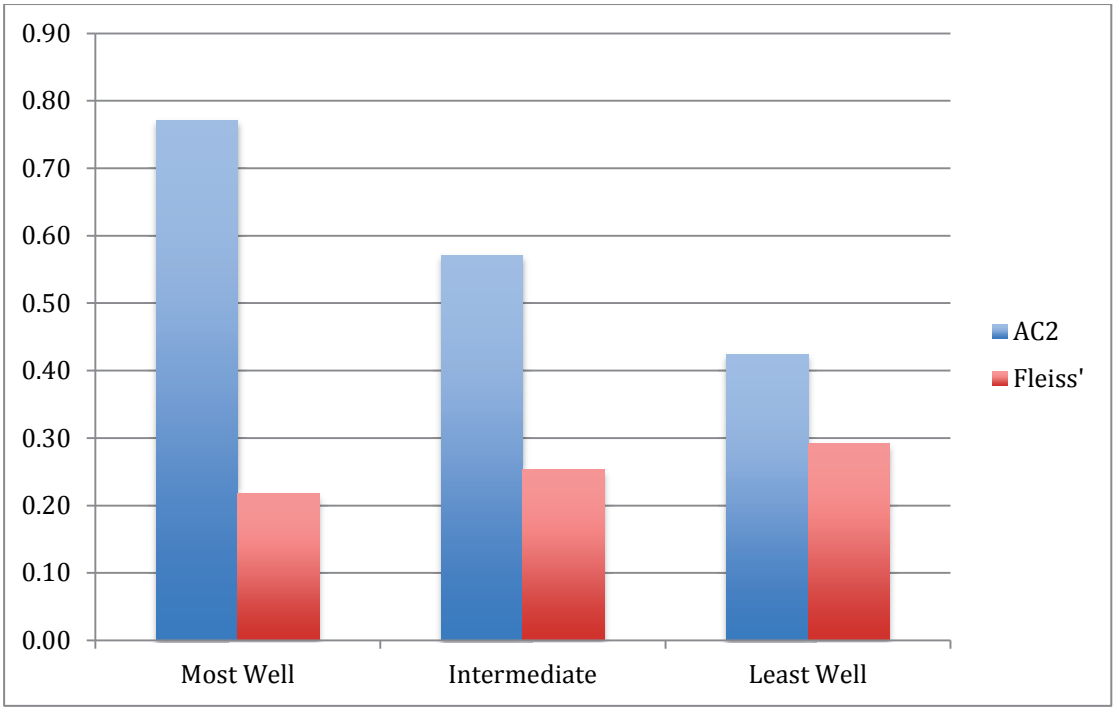

Figure 4.5.1 Average agreement of three wellness groups using Linearly weighted Gwet's AC2 and Fleiss' Kappa

Table 4.5.2 and Figure 4.5.1 clearly show the vast difference in outcomes produced by the two statistics with the use of Fleiss' Kappa on data with the wrong marginality. Table 4.5.2 shows that Fair agreement occurs among the raters in the wellness groups using Fleiss' Kappa. This is in strong contrast with the Moderate to Substantial agreement derived with Gwet's AC2 statistic. What is noticeable is the vast difference between the value of the simple agreement and the Fleiss' Kappa results, especially in the Most Well group. Here the Fleiss' Kappa is 0.22 ±0.03, whereas the simple agreement was 0.85 ±0.01; a very large reduction apparently caused for removing agreement by chance. In contrast, the AC2 value is 0.77 ±0.02, which seems a much more realistic chance-removed value.

The presentation of the data in Figure 4.5.1 clearly illustrates the differences between the AC2 and Fleiss' Kappa results. The Most Well group had Slight Fleiss' weighted Kappa agreement, a misleading outcome when compared with the AC2 statistic, which correctly evaluated the chance removed agreement as Substantial. The preponderance of zeroes selected for Descriptors defining patients' health of subjects in the Most Well group, causes these data to be free marginal; thereby making the use of Fleiss' Kappa totally inappropriate.

In contrast, in the Least Well group, the convergence of the two statistical measures indicated these data were becoming more fixed marginal. Both were defined as Fair, albeit at the two extremes of Landis' classification.

Since no published reliability studies[24, 25, 99], which used the Kappa statistics, indicate the marginality of the data collected, there is no way to determine how reliable the resulting agreements really are. Unless these data were absolutely fixed marginal, the level of agreement estimated with Fleiss' Kappa would always be lower than it actually is. Due to a lack of information relating to the marginality the reduction in agreement cannot be determined. This again confirms that Fleiss' Kappa statistic should not be used to determine agreement in data that is not or cannot clearly be determined as fixed marginal.

4.6 Agreement in the Individual Descriptors

The sources of inter-rater discord was also investigated by estimating the Descriptors were next processed individually; comparing the scores from the five practitioners with AC2 linearly weighted inter-rater statistics, to explore inter-rater agreement in each of the Descriptors. It is important to look at the inter-rater agreement in each individual Descriptor. Which Descriptors performed most adequately, and which did not? This investigation should provide interesting insights into which descriptors that need further improvement to optimise inter-rater agreement.

The 16 descriptors of the DSOM format are presented below in Table 4.6.1. Percentage agreement and AC2 values and the standard errors are listed. The Descriptors are sorted from lowest to highest according to the AC2 value obtained from the sores provided by raters,. The two columns on the right represent the average TPS and percentage score selections previously reported and discussed in section 4.3.1, to highlight how heavily each Descriptor was utilised.

Table 4.6.1 Individual Descriptor Agreement expressed in linearly weighted percentage and AC2, and the Average TPS and frequency of occasions the descriptor was scored one or greater.

Descriptor	Simple Agreement	AC2	Average TPS	Frequency of selection %
Qi Stag	0.70 ±0.02	0.29 ±0.05	1.10	67
Spleen	0.70 ±0.02	0.29 ±0.05	1.41	67
Liver	0.71 ±0.02	0.32 ±0.06	0.89	65
Qi Xu	0.72 ±0.02	0.35 ±0.06	1.37	66
Kidney	0.72 ±0.02	0.42 ±0.06	0.99	51
Heat	0.76 ±0.02	0.53 ±0.06	0.67	48
Yin Xu	0.76 ±0.03	0.59 ±0.06	0.69	38
Blood Stag	0.77 ±0.02	0.62 ±0.06	0.57	35
Damp	0.80 ±0.03	0.65 ±0.06	0.46	37
Heart	0.80 ±0.03	0.68 ±0.07	0.21	33
Blood Xu	0.80 ±0.03	0.69 ±0.06	0.31	29
Yang Xu	0.82 ±0.03	0.74 ±0.05	0.36	26
Lung	0.85 ±0.02	0.76 ±0.05	0.33	30
Cold	0.84 ±0.03	0.77 ±0.05	0.21	23
Phlegm	0.86 ±0.02	0.81 ±0.04	0.11	19
Dry	0.86 ±0.03	0.82 ±0.04	0.20	20
All Descriptors	0.78 ±0.03	0.60 ±0.02	0.62	41

What is troubling is the low AC2 agreement found in the descriptors that were used most frequently and also had the highest TPS averages. The ‘top five’

Descriptors used all performed the worst in terms of inter-rater agreement, with the four most often selected (being Qi Stagnation, Spleen, Liver and Qi Xu) achieving only Fair agreement, and the fifth (Kidney) reaching Moderate agreement on Landis and Kochs' scale. The differences between percentage agreement and AC2 values are especially striking in these descriptors. At the other end of Descriptor selection Phlegm and Dryness had Almost Perfect chance-removed agreement. The agreement in these least selected Descriptors however was made up of a preponderance of zero selections.

Agreement across all Descriptors and all subjects is a Moderate agreement on the basis of AC2 value of 0.60 ±0.02. However, the observations that the most frequently used Descriptors that define a subject's CM health AND the overall reduced agreement in the Least Well group is interesting and should be investigated to determine if this outcome can be changed by some appropriate interventions. As a consequence, the scoring pattern in each individual Descriptor, the total scores of rater-Descriptor combinations and the average score from all raters in each Descriptor are presented in Tables 4.6.2 (a b and c) in an attempt at understanding the reason for these problematic outcomes..

Table 4.6.2 (a) Day One Descriptor Raw Scores and Average Scores

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Average
Cold	20	3	10	6	17	11.2
Heat	39	27	18	25	57	33.2
Damp	22	23	11	25	37	23.6
Dry	13	4	18	0	16	10.2
Phlegm	18	0	12	13	12	11
Qi Xu	25	37	42	22	20	29.2
Blood Xu	18	3	8	11	28	13.6
Qi Stag	36	40	47	26	43	38.4
Blood Stag	1	26	15	11	31	16.8
Yin Xu	29	9	10	26	40	22.8
Yang Xu	17	4	19	9	13	12.4
Liver	51	24	49	42	44	42
Heart	10	10	14	8	20	12.4
Spleen	40	34	35	22	54	37
Kidney	19	19	35	25	35	26.6
Lung	15	12	25	12	24	17.6

Table 4.6.2.1 (b) Day Two Descriptor Raw Scores and Average Scores

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Average
Cold	15	3	6	0	6	6
Heat	21	6	5	10	6	9.6
Damp	7	7	11	10	17	10.4
Dry	25	29	25	47	44	34
Phlegm	11	4	5	20	12	10.4
Qi Xu	25	29	25	47	44	34
Blood Xu	9	4	2	19	11	9
Qi Stag	19	27	28	20	32	25.2
Blood Stag	0	14	15	2	15	9.2
Yin Xu	23	5	4	8	11	10.2
Yang Xu	13	0	11	18	5	9.4
Liver	20	22	21	29	31	24.6
Heart	17	3	9	17	13	11.8
Spleen	19	18	23	30	44	26.8
Kidney	23	18	14	30	34	23.8
Lung	11	7	7	19	17	12.2

Table 4.6.2.1 (c) Day Three Descriptor Raw Scores and Average Scores

Day Three	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Average
Cold	15	7	0	5	3	6
Heat	12	16	10	14	15	13.4
Damp	8	4	3	3	9	5.4
Dry	10	1	9	9	6	7
Phlegm	12	0	0	0	0	2.4
Qi Xu	21	7	15	17	16	15.2
Blood Xu	11	3	7	13	9	8.6
Qi Stag	27	1	19	18	19	16.8
Blood Stag	20	3	5	14	17	11.8
Yin Xu	5	10	22	16	14	13.4
Yang Xu	14	6	2	7	0	5.8
Liver	23	10	22	14	25	18.8
Heart	14	16	8	24	13	15
Spleen	27	6	15	16	15	15.8
Kidney	17	27	24	20	17	21
Lung	14	1	1	4	0	4

Looking at these data, significantly larger differences in the total scores were apparent in some Descriptors than in the total of all Descriptors combined.

There were twelve zero scored rater-Descriptor combinations. On Day one there were two zero scored rater-Descriptor combinations, namely, Dry and Phlegm, while on day two there were three, one each in Cold, Yang Xu and Blood Stagnation and on Day three there were seven zero scored rater-Descriptor combinations. The highest number of zero scored rater-Descriptor combinations occurred on days three, which was understandable as there were only eight subjects interviewed that day which is half the number of the other two days. On day three, Phlegm was the least frequently scored Descriptor with four of the five raters scoring it as zero for all eight subjects. The other three zero scored Descriptors that occurred on day three were

Cold, Yang Xu and Lung. The non-scoring of rater-Descriptor combinations will naturally lead to high inter-rater agreement. It can be observed that all the Descriptors that had zero rater totals also had the five highest chance-removed agreements.

Another important observation is the large differences in the total scores that sometimes occurred between raters scoring the same Descriptor. One striking example is Blood Stagnation on day one; rater 5 scored 31, rater 2 scored 26, rater 3 15, rater 4 gave 11, but rater 1 scored a total of only 1! Qi Stagnation on day three is another example; rater 1 scored a total of 27, raters 3 and 5 gave a total of 19 and rater 4 scored 18, in contrast, rater 2 gave a total of a single point only. Low scoring by some raters in certain Descriptors while other raters score strongly must exert a downward effect on agreement.

The agreements presented in Table 4.6.1 are for all days, while Tables 4.6.2 (a-c) are the separate days score choices. The large variation in scores in certain Descriptors on certain days does not however explain the low agreement estimated in the most used descriptors. The ‘big five’ i.e Qi Xu, Qi Stagnation, Liver, Spleen and Kidneys had the highest non-zero selections from all raters on all days as summarised in the final column of Table 4.6.1, but the lowest chance-removed AC2 agreement presented in column three. While no zero total scores occurred in these rater-Descriptor combinations, in some cases the differences were still large, with the total score of one rater often being double or more of another, such as Qi Stagnation on day three, while on day 1 raters 1 and 2 gave total scores of 51 and 24 respectively for

Liver and on day the total given by two rater for spleen was 18 less than half 44 scored by 5.

4.7 Individual Descriptor Agreement and Total Descriptor Score

Due to the lower than average AC2 agreements in the least well group and the most used descriptors, an investigation was undertaken, to understand the dynamics of agreement observed. Here, each the data for individual Descriptors was decoupled from the subjects and the data sorted into Intra-Descriptor Wellness Groups. Since all previous work described in this thesis was done with each patient's descriptors in combination and sorted into groups according to the subjects' TPS, this approach is a radical departure from all previous attempts mentioned above. In this case a new measure, called the Total Descriptor Score (TDS), was introduced to allow the ranking of subjects in each descriptor. The TDS for a subject was calculated by addinf the score of the five rates adding for each Descriptor. Similarly to the TPS, the TDS was used to sort the data of each Descriptor into three groups according to the TDS.

Therefore, the total of all TDS attributed to each subject in each Descriptor was determined and then based upon these scores the subjects scores in each Descriptor was sorted into three groups; Least Well, Intermediate and Most Well. Each group consisted of lists of Descriptor scores from the five practitioners, and each descriptor wellness group was unique. The criteria for inclusion into the Least Well group in a given Descriptor was an average TDS

of three or greater, while the Intermediate group contained subjects whose average TDS was greater than one and less than three. The Most Well group consisted of subjects whose average score was one or less. If there were sufficient high scoring data available, the Least Well group would have included only subjects with an average TPS that was four instead of three.

The aim of this investigation was to understand how different average TDS scores affected inter-rater agreement in each individual descriptor. Upon reflection it was obvious that agreement must be perfect at the two extremes, where all raters scored a patient in an individual Descriptor as zero, or all score a descriptor maximally. Cases where the average TDS is lower than one, or greater than 3 both approach this scenario. The Most Well groups and the least used Descriptors already reported in section 4.5 demonstrate the agreement results obtained in the extreme of low scoring. To investigate the effects of higher and intermediate scoring selection, the data in each descriptor was sorted according to the average TDS and then split into the three groups and the linearly weighted percentage agreement and AC2 were calculated.

Unlike in the Most Well and Least Well individual Descriptor groups, agreement in the Intermediate cluster has the potential to be poor. Each subject may attract mixed scores, with the possibility that some raters score the descriptor zero and others highly, even maximally, which will then engender low agreement especially in the case of chance-removed

agreement. Two tables were therefore assembled, simple agreement in 4.7.2 and the AC2 statistic in 4.7.3.

Table 4.7.1 Number of subjects that were included in the intra-descriptor subgroups

	Least Well	Intermediate	Most Well
Blood Stag	-	15	27
Blood Xu	-	11	31
Cold	-	8	34
Damp	1	15	26
Dry	-	8	34
Heart	3	11	28
Heat	5	18	19
Kidney	10	13	19
Liver	12	20	10
Lung	4	9	29
Phlegm	-	7	35
Qi Stag	9	24	9
Qi Xu	9	22	11
Spleen	9	20	13
Yang Xu	1	9	32
Yin Xu	3	13	26
Average	4	14	24

Many Descriptors in Least Well group did not have sufficient raters scoring at the required level, with eight or 50% of the sixteen having one or less. The average number of subjects who fulfilled the criteria for inclusion in the Least Well groups were only four. The threshold for calculating inter-rater agreement in an individual Descriptor group was set at three subjects or more.

The linearly weighted simple and AC2 agreements in each Descriptor and the intra-descriptor subgroups are next presented in Tables 4.7.2 and 4.7.3.

Table 4.7.2 Linearly weighted simple agreement in the intra-descriptor subgroups

Groups	Least Well		Intermediate		Most Well	
Statistic	Simple Agreement	Std Error	Simple Agreement	Std Error	Simple Agreement	Std Error
Blood Stag	N/A		0.64	0.02	0.85	0.03
Blood Xu	N/A		0.58	0.03	0.88	0.02
Cold	N/A		0.59	0.03	0.89	0.02
Damp	N/A		0.63	0.03	0.89	0.02
Dry	N/A		0.61	0.02	0.92	0.02
Heart	0.73	0.04	0.57	0.03	0.90	0.02
Heat	0.77	0.07	0.65	0.02	0.87	0.03
Kidney	0.74	0.04	0.56	0.02	0.81	0.03
Liver	0.73	0.04	0.2	0.03	0.87	0.03
Lung	0.78	0.06	0.64	0.02	0.92	0.02
Phlegm	N/A		0.57	0.03	0.91	0.02
Qi Stag	0.76	0.05	0.63	0.01	0.82	0.03
Qi Xu	0.75	0.03	0.66	0.02	0.84	0.03
Spleen	0.84	0.03	0.66	0.02	0.79	0.03
Yang Xu	N/A		0.57	0.03	0.90	0.02
Yin Xu	0.79	0.07	0.59	0.03	0.85	0.02
Average	0.76	0.05	0.58	0.02	0.87	0.02

Table 4.7.3 Linearly weighted AC2 agreement in the intra-descriptor subgroups

Groups	Least Well		Intermediate		Most Well	
Statistic	AC2	Std Error	AC2	Std Error	AC2	Std Error
Blood Stag	N/A		0.17	0.05	0.80	0.04
Blood Xu	N/A		0.03	0.06	0.85	0.04
Cold	N/A		0.14	0.09	0.87	0.03
Damp	N/A		0.17	0.06	0.87	0.03
Dry	N/A		0.20	0.03	0.91	0.03
Heart	0.43	0.13	-0.03	0.07	0.88	0.03
Heat	0.53	0.17	0.20	0.04	0.82	0.02
Kidney	0.46	0.11	0.05	0.05	0.73	0.05
Liver	0.41	0.10	0.05	0.04	0.82	0.06
Lung	0.58	0.15	0.24	0.04	0.90	0.02
Phlegm	N/A		0.24	0.05	0.90	0.02
Qi Stag	0.49	0.14	0.14	0.04	0.72	0.06
Qi Xu	0.5	0.07	0.19	0.06	0.75	0.06
Spleen	0.26	0.06	0.19	0.06	0.64	0.07
Yang Xu	N/A		0.07	0.06	0.87	0.03
Yin Xu	0.57	0.20	0.09	0.06	0.80	0.04

While the subject Descriptor requirements for inclusion into the Most Well and Least Well Descriptor groups literally forced consistency in scoring, the consistency of the score in the Intermediate group had the potential to vary wildly; poor consistency would be reflected in a low value in a group's chance-removed inter-rater agreement.

The Intermediate Descriptor groups performed worst of all, with agreement in all Descriptors in this group being Poor, while the Least Well Descriptor groups agreement rated mostly Moderate and on one occasion as Slight. This is in stark contrast to the Almost Perfect agreement obtained in the Most Well Descriptor groups. It seemed that cause for the progressively worsening

agreement from the Most Well to the Intermediate and especially in Least Well groups presented and discussed in 4.4 was the due to the increasing presence of Intermediately scored descriptors in the included subjects.

The standard errors reported were high in certain instances due to low sample sizes.

4.8 DSOM Questionnaire Agreement with the Practitioner Diagnoses

To analyse the potential representation of practitioners' diagnoses with the DSOM questionnaire, the results from the DSOM questionnaire data were sent for processing to Lee in South Korea and the diagnoses made with the undisclosed algorithm were received in due course. The data sent to Lee was without any diagnoses made from the five practitioners, thereby 'blinding' the questionnaire data from the diagnoses made at the UTS clinic.

By comparing the levels of agreement attained with and without the inclusion of the DSOM questionnaire diagnoses, it was possible to determine if the questionnaire-derived diagnoses were representative of what the practitioners were diagnosing. If the agreement improved or was sustained, then the DSOM-derived diagnoses could be said to be representative. If the level of agreement declined, then the DSOM diagnoses were not representative.

There are three metrics returned by the DSOM, raw scores between 0-100, a weighted score between 0-10 and a reliability index that consisted of 4 categories, LL, LH, HL and HH. It was decided to use the weighted score to compare to the practitioner data. As the weighted score ranged between 0-10 and the practitioner data varied between 0-5, the weighted scores were reduced by 50% to allow adequate comparison with the practitioner data.

The data from table 4.4.1, which reports the agreement between the five practitioners using the DSOM format, is reproduced below.

Table 4.8.1 Agreements in each Wellness group

Groups	Weighted Percentage	AC2	TPS Average
Most Well	85 ±0.01	0.77 ±0.02	49
Intermediate	76 ±0.01	0.57 ±0.03	93
Least Well	73 ±0.01	0.42 ±0.03	132
All Groups	78 ±0.01	0.60 ±0.02	91

Table 4.8.2 Agreements between the five practitioners with DSOM questionnaire data

Groups	Weighted Percentage	AC2	TPS Average
Most Well	86 ±0.01	0.80 ±0.02	49
Intermediate	76 ±0.01	0.61 ±0.03	93
Least Well	71 ±0.01	0.44 ±0.03	132
All Groups	78 ±0.01	0.63 ±0.02	91

Table 4.8.3 Changes in agreements between the five practitioners after the inclusion of DSOM questionnaire data.

Groups	Weighted Percentage	AC2	TPS Average
Most Well	1 ±0.01	0.03 ±0.03	49
Intermediate	0 ±0.01	0.04 ±0.04	93
Least Well	-2 ±0.01	0.01 ±0.04	132
All Groups	0 ±0.01	0.03 ±0.03	91

The levels of agreement did indeed seem to improve, specifically within the AC2 linearly weighted agreement as shown in table 4.8.3. However, these improvements generally occurred within the square root of the sum of squares of the standard errors, so are not substantial. For the difference between the two results to be of significance, the scores must differ by a greater amount than the combined standard errors of each result.

While the fact that the DSOM questionnaire results seemed representative of what five practitioners would select as a diagnosis using the descriptors available within the DSOM is interesting, the problems with the questionnaire previously outlined in 2.7 still provide sufficient disincentive for its use as a diagnostic tool. A question needed to be asked, whether the Descriptors utilised within the DSOM are indeed the appropriate factors that should be available for scoring in such a diagnostic format. Perhaps other Descriptors should be included, or maybe some of the Descriptors present in the DSOM were not actually needed, such as Dryness and Phlegm?

4.9 Discussion of results of DSOM data collection

It appears that no previous experiment has been performed on subjects whose health status is unknown, that is from an open population. Further, the subjects were diagnosed by a significant number, namely five, appropriately experienced CM practitioners. This is a much larger number of raters than in previous studies found in the literature, so that reliable statistical results could be obtained. Finally, the inter-rater agreement was calculated with the linearly weighted, chance-removed AC2 statistic. Substantial, 0.60 ±0.02, inter-rater agreement was obtained in this experiment in which the DSOM format was used to record diagnoses. Since this level of agreement is higher that those previously mentioned in the literature, the Substantial agreement obtained is an encouraging intitial result and could indicate that results obtained in previous studies could have been underestimating the level of inter-rater agreement.

Once the subjects had been sorted into three Wellness groups according to the TPS for each subject, the inter-rater agreement between the practitioners was calculated for each of the Wellnes groups once again using the linearly weighted AC2 statistic. Substantial chance-removed agreement of 0.77 ±0.02 was calculated in the Most Well Group; Moderate agreements of 0.57 ±0.03 and 0.42 ±0.03 were estimated in the Intermediate and the Least Well Groups. The significant decline in inter-rater agreement from the Most Well Groups to the Intermediate and the continuing decline in agreement Least Well groups is observed.

It has to be remembered that the practitioners were dealing with subjects drawn from an open population, of whom they had no other knowledge than that established in the diagnostic interview. Unlike other research studies into inter-rater diagnostic agreement, which normally deal with specifically determined diagnostic categories, the whole gamut of possible disease or wellness states was open. It therefore appears that practitioners can competently identify subjects after one interview those who are healthy, however, as the subjects' health declines leading to increased diagnostic complexity, the inter-rater agreement declines. Indeed, the very subjects that arguably are most in need of effective diagnoses and therefore treatment are the most disadvantaged.

This is a very important result, which has not been mentioned in the literature. In order to investigate the possible causes of this phenomenon individual Descriptor agreement was next explored. Not surprisingly therefore, the lowest chance-removed agreement occurred in the most frequently scored Descriptors; with agreement in the five most highly scored Descriptors averaging Fair inter-rater agreement of 0.33 ±0.06. The top five were Liver, Spleen, Kidney Qi Deficiency and Qi Stagnation which also are the most frequently used Descriptors in the UTS clinic. This means that the most common descriptors are also the ones about which there is most diagreement amongst raters, so that there seems to be a failure in the education process if five raters cannot agree on this most fundamental descriptors.

A final analysis of agreement within individual Descriptors was conducted. Each Descriptor's data set was sorted into intra-descriptor wellness groups based upon the average TPS of each descriptor. In the Most Well intra-descriptor groups, defined by an average TPS score of one or less, an Almost Perfect chance-removed average agreement of 0.82 ±0.04 was calculated across all Descriptors. In the Least Well intra-descriptor groups, with an average TPS of three or greater in each Descriptor, a Moderate average agreement of 0.52 ±0.12 was calculated.

In the Intermediate group, defined by an average individual Descriptor TPS of above one and below three, the agreement was Slight, with an average 0.13 ±0.05 across all Descriptors. Based upon the observations made in the wellness groups when all Descriptors were included, one would have expected that the agreement of the Intermediate intra-descriptor groups to be between that calculated in the Most Well and Least Well.

This outcome is counter-intuitive when the previous results are considered. This leads to a new understanding that agreement is lowest where raters score descriptors in an intermediate fashion and higher when descriptors are scored more heavily.

The poor agreement in the Least Well Group of all descriptors is due to a significant proportion of intermediately classified intra-descriptors occurring therefore pushing down the agreement. This should be a vital insight that could be used to target areas of poorest performing agreement. This seems to

be the first time inter-rater agreement has been investigated in an open population using a format that enables comparitively no restriction in choice of diagnosis to that found in the literature review. This is exciting, as identifying problems is the first step towards their solution, and now there is a target for improvement in this critical area in future studies.

Chapter 5 UTS Outpatient Clinic Data

To investigate whether the descriptors used by the DSOM were appropriate choices for inclusion within a diagnostic format, a large dataset of diagnoses from a practising clinic had to be examined.

A database that recorded acupuncture diagnoses was investigated to enquire as to which Diagnostic Patterns (DPs) were typically selected in CM acupuncture practice in Australia. Over 60,000 patient diagnostic records at UTS CM outpatient clinic over a twelve-year period[100] were analysed to form a long list of DPs.

The UTS clinical data are an amalgamation of a large number of patients' records: reflective of many practitioners, and recorded 109 unique patterns. These data might be seen as being typical of a CM clinical setting in an open population. At the UTS outpatient clinic, no limit to the number of DPs that could be allocated to a patient was made. The first three diagnoses however, made up the vast bulk of DPs selected, as is shown next in table 5.0.1

Table 5.0.1 Percentage of total diagnoses according to order of choice

Diagnoses	Count	Ratio to all diagnoses	Aggregate
Primary	60228	59%	59%
Second	33713	33%	92%
Third	5842	6%	98%
Fourth	1759	2%	99%
Fifth	554	1%	100%
Sixth	122	0%	100%
Seventh	24	0%	100%
Eighth	8	0%	100%
Ninth	2	0%	100%
Tenth	1	0%	100%
	102253	100%	

The aggregate percentage of selections in the fourth column of table 5.0.2 shows that the first, second and third DPs allocated by the practitioners represent 98% of all DPs selected. The fact that 98% of all diagnoses given to the patients at the clinic consisted of the first three diagnoses prompted a decision to look only at these first three selections.

Next, an examination of how many patterns were frequently used was made. As shown next in table 5.0.2, the first 56 patterns made up approximately 95% of practitioners' DP selections in the first, second or third choices by the acupuncture practitioners in the UTS Chinese medicine outpatient clinic.

Table 5.0.2 Practitioners top 56 Diagnostic Pattern selections as a percentage of all Diagnostic Patterns

Diagnoses	Top 56	All DPs	Ratio of Top 56 to All DPs
Primary	57420	60228	95%
Second	31577	33713	94%
Third	5590	5842	96%

Based upon the information presented in tables 5.0.1 and 5.0.2 the decision was made to examine the 56 patterns selected as first choice, second or third choices, to determine the most common 56 patterns chosen, which are next presented in Table 5.0.3. Although the selected DPs represented just over half of the diagnostic choices available, they constituted 93% of all diagnoses selected.

Table 5.0.3 The top 56 patterns selected at UTS outpatient clinic

Pattern	Count	Percent
Spleen Qi Xu	11671	11.41%
Liver Qi Stagnation	11089	10.84%
Qi & Blood Stagnation	7827	7.65%
Qi & Blood Stagnation in the Bladder Channel	6084	5.95%
Kidney Qi Xu	5177	5.06%
Kidney Yin Xu	5056	4.94%
Qi & Blood Stagnation in the Gall Bladder Channel	4465	4.37%
Bi Syndrome	2963	2.90%
Lung Qi Xu	2432	2.38%
Qi Stagnation in the Bladder Channel	2255	2.21%
Qi Stagnation in the Gall Bladder Channel	2244	2.19%
Wind Heat Attacks the Lung	1763	1.72%
Wood (liver) invades Earth (spleen)	1614	1.58%
Qi Stagnation (localised trauma)	1492	1.46%
Liver Yin Xu	1466	1.43%

Table 5.0.3 The top 56 patterns selected at UTS outpatient clinic (continued)

Pattern	Count	Percent
Blood Xu	1427	1.40%
Qi & Blood stagnation in the Large Intestine Channel	1392	1.36%
Qi & Blood Stagnation in the Small Intestine Channel	1370	1.34%
Damp Heat in the Spleen	1316	1.29%
Kidney Yang Xu	1130	1.11%
Liver Blood Xu	986	0.96%
Wind Cold Attacks the Lung	971	0.95%
Phlegm Heat Obstructing the Lung	886	0.87%
Damp Heat in the Gall Bladder	880	0.86%
Liver Yang Rising	874	0.85%
Heart Qi Xu	872	0.85%
Cold Damp in the Spleen	750	0.73%
Qi Stagnation in the Small Intestine Channel	733	0.72%
Qi Stagnation in the Large Intestine Channel	720	0.70%
Damp Heat in the Liver	674	0.66%
Kidney Jing Xu	669	0.65%
Liver Heat Rising	599	0.59%
Heat in the Blood	575	0.56%
Qi & Blood Stagnation in the Stomach Channel	548	0.54%
Liver Wind (Internal - moving)	535	0.52%
Qi & Blood Stagnation in the Liver Channel	535	0.52%
Heart Yin Xu	509	0.50%
Cold Stagnation in the Gall Bladder Channel	485	0.47%

Table 5.0.3 The top 56 patterns selected at UTS outpatient clinic (continued)

Pattern	Count	Percent
Qi Stagnation in the Liver Channel	479	0.47%
Cold Stagnation in the Bladder Channel	436	0.43%
Lung Yin Xu	436	0.43%
Liver Fire (blazes)	414	0.40%
Qi Stagnation in the Stomach Channel	414	0.40%
Qi & Blood Stagnation in the Triple Energiser Channel	410	0.40%
Heart Fire (blazes upwards)	400	0.39%
Damp Heat in the Large Intestine	399	0.39%
Damp Heat in the Bladder	390	0.38%
Stomach Heat	376	0.37%
Qi & Blood Stagnation in the Spleen Channel	374	0.37%
Spleen Blood xu	348	0.34%
Spleen Yang Xu	347	0.34%
Heat in the Stomach	336	0.33%
Heat in the Heart	317	0.31%
Heart and Kidney not Communicating	292	0.29%
Phlegm Cold Obstructing the Lung	176	0.17%
Qi & Blood Stagnation in the Conception Vessel (Ren Mai)	150	0.15%
The last 53 patterns	8725	8.53%
All Patterns	102253	100%

The data presented in table 5.0.3 indicate the top 56 patterns selected as first choice, and when these patterns were also selected as second or third choices.

As there were 109 patterns recorded in the UTS database, another 53 patterns were rarely used, representing approximately 9% of all diagnoses given. Each of these remaining 53 patterns was selected at or below 0.15% occasions each.

Using the 56 patterns just presented, the likelihood of different practitioners choosing exactly the same primary pattern is extremely improbable. How much more difficult would be the attainment of exactly the same combination of two, three or even more patterns?

5.1 UTS Student Clinic Data Mapped to the DSOM Descriptors

An attempt to map the 56 patterns most frequently used at the UTS clinic with the DSOM Descriptors was next made. The 93,528 patterns recorded at the clinic translated to a total of 205,582 descriptor selections. The number of Descriptors required was mostly two and in some cases three Descriptors that required to map a DP. The result of this mapping attempt is presented next in table 5.1.1

Table 5.1.1 Mapping of Descriptors from the most popular 56 patterns selected at UTS CM Outpatient Clinic

Descriptor	Number of selections	Percentage of selection
Qi Stag	42102	20.6%
Liver	26860	13.1%
Xue Stag	24076	11.8%
Kidney	21053	10.3%
Qi Xu	20152	9.9%
Spleen	18530	9.1%
Heat	9325	4.6%
Lung	9175	4.5%
Yin Xu	9010	4.4%
Heart	4903	2.4%
Damp	5471	2.2%
Xue Xu	3336	1.6%
Cold	2818	1.4%
Yang Xu	1477	0.7%
Phlegm	1062	0.5%
Dryness	0	0.0%
Wind	3269	1.6%
Bi	2963	1.4%
Total	205582	100%

There are two observations; the first is the very slight (0.5%) requirement of the Descriptor phlegm and the total absence of dryness within the 56 patterns commonly selected at the UTS outpatient clinic. The second is the presence of the additional two diagnostic Descriptors needed to map the patterns, Wind and Bi Syndromes. Wind and Bi added up to just over 3% of total Descriptor selection.

Bi syndrome was used at the UTS as a separate pattern and was never used in conjunction with other Descriptors. It seemed only used to describe pain.

5.2 DSOM Study and UTS Outpatient Clinic Data Compared

The Descriptors selected within the DSOM data were next compared to the UTS outpatient clinic data mapped to the DSOM, to determine whether the data from both origins had any similarities. The data was sorted according to DSOM Descriptor selections and are next presented in table 5.2.1.

Table 5.2.1 Descriptor choices in the DSOM and UTS outpatient clinic data

Descriptor	DSOM Data	UTS Clinic	Rank Difference
Qi Stag	1	1	0
Spleen	1	6	5
Qi Xu	3	5	2
Liver	4	2	2
Kidney	5	4	1
Heat	6	7	1

Table 5.2.1 Descriptor choices in the DSOM and UTS outpatient clinic data (continued).

Descriptor	DSOM Data	UTS Clinic	Rank Difference
Yin Xu	7	9	2
Damp	8	11	3
Blood Stag	9	3	6
Heart	10	10	0
Lung	11	8	3
Blood Xu	12	12	0
Yang Xu	13	14	1
Cold	14	13	1
Dry	15	16	1
Phlegm	16	15	1

The two studies when mapped to the DSOM Descriptors correlated at 0.86 and suggest a significant linear relationship. Generally, a correlation co-efficient above ±0.50 is considered strong. The first six choices in both data are the same except for their order. In both datasets, dryness and phlegm were least often chosen.

Notable differences occur between the rankings in Blood Stagnation and Spleen. Blood Stagnation, with the greatest ranking difference of six places is positioned within the UTS data as the third choice, but is the ninth in the DSOM data, and Spleen, which has a five-point rank difference.

The difference between the Blood stagnation rankings in the two datasets may be explained by the fact that subjects attending the UTS outpatient clinic are often in attendance for acute or chronic pain, which frequently includes a diagnosis of blood stasis. The subjects within the DSOM data on the other hand, were recruited from an open population from the community that included many well subjects, and treatment was not offered.

Chapter 6 The Chinese Medicine Diagnostic Descriptor

After reviewing the UTS student and teacher clinic data[100] in chapter 5, the Chinese Medicine Diagnostic Descriptor (CMDD) was developed by editing the descriptors of the DSOM to accommodate all patterns reported in Australian clinical conditions. The alterations made extend the DSOM system from a format that was initially designed to specify female reproductive health from a CM perspective, to one that describes a patient's overall CM constitutional health in both sexes and in all circumstances.

Whilst there had been 16 Descriptors in the original DSOM format, 14 were retained, one deleted, two merged, and one added to arrive at the 15 of the CMDD. All changes between the two formats took place in the Disease Cause (bing yin) category. Phlegm and Damp were merged to Damp. Dryness was removed, as it was not present in any pattern recorded at the UTS Chinese medicine outpatient clinic[100]. Jing was not mentioned in the DSOM, and is seen as a presentation of Yin in the CMDD.

Bi Syndromes were not recognised in the DSOM. This approach was preserved in the CMDD and Bi Syndromes are seen as Cold, Qi or Blood Stasis depending on their cause. As in the DSOM, issues in the yin and yang organs are recorded with a five-phase approach in the CMDD. This leads to yang organ and channel problems in all cases being attributed to the yin organ parent. This is consistent with the CMDD's objective of recording a constitutional picture.

Finally, Wind was added as a Descriptor to the CMDD to complete an adequate list of Descriptors to allow representation of all the patterns found in the UTS Chinese medical outpatient clinic database. It is understandable that Wind would not be included as a Descriptor in the DSOM as this Disease Cause does not usually affect women's reproductive health.

The CMDD is presented in the figure below, with the Descriptors arranged in three columns, the first pathogenic factors, the second substances and the last the yin organs. Scores allocated to these factors are used to describe a patient's condition.

Figure 6.0.1 The Chinese Medicine Diagnostic Descriptor

Descriptor Definitions

The fifteen Descriptors of the CMDD are defined using the appropriate definitions contained the WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region[70]. Quoting directly from this reference each of the Descriptors is now briefly summarised.

Heart: "the organ located in the thoracic cavity above the diaphragm, which controls blood circulation and mental activities";

Spleen: "the organ located in the middle energizer below the diaphragm, whose main function is to transport and transform food, upbear the clear substances, keep the blood flowing within the vessels, and is closely related to the limbs and flesh";

Lung: "a pair of organs located in the thoracic cavity above the diaphragm, which control respiration, dominate qi, govern diffusion and depurative down-bearing, regulate the waterways, and are closely related to the function of the nose and skin surface";

Kidneys: "a pair of organs located in the lumbar region, which store vital essence, promote growth, development, reproduction, and urinary function, and also have a direct effect on the condition of the bone and marrow, activities of the brain, hearing and inspiratory function of the respiratory system";

Liver: “the organ located in the right hypochondrium below the diaphragm, which stores blood, facilitates the coursing of qi, and is closely related to the function of the sinews and eyes”;

Qi Deficiency: “a general term for deficiency of qi that leads to decreased visceral functions and lowered body resistance”;

Blood Deficiency: “any pathological change characterized by deficiency of blood, which fails to nourish organs, tissues and meridians/channels”;

Qi Stagnation: “a pathological change characterized by impeded circulation of qi that leads to stagnation of qi movement and functional disorder of organs, manifested as distention or pain in the affected part”;

Blood Stagnation: “a pathological product of blood stagnation, including extravasated blood and the blood circulating sluggishly or blood congested in a viscus, all of which may turn into pathogenic factor, the same as blood stasis or stagnant blood”;

Yin Deficiency: “a pathological change marked by deficiency of yin with diminished moistening, calming, down bearing and yang-inhibiting function, leading to relative hyperactivity of yang qi”;

Yang Deficiency: “a pathological state characterized by deficiency of body’s yang qi that leads to diminished functions, decreased metabolic activities, reduced body reactions as well as deficiency-cold manifestations”;

Dampness: “a pathogenic factor characterized by its impediment to qi movement and its turbidity, heaviness, stickiness and downward flowing properties, also called pathogenic dampness”, which also includes **Internal Dampness** “produced in the body due to yang deficiency of the spleen and kidney with decreased fluid transportation and transformation and resultant water stagnation” and **Phlegm**: “1) pathologic secretions of the diseased respiratory tract, which is known as sputum; 2) the viscous turbid pathological product that can accumulate in the body, causing a variety of diseases”;

Cold: (external) “as one of the six excesses that causes external cold pattern/syndrome and (internal) cold in the interior due to deficiency of yang qi or preponderance of yin cold” or **Cold** “as a pathogenic factor characterized by the damage to yang qi, deceleration of activity, congealing and contracting actions, also called pathogenic cold”;

Heat: “as a pathogenic factor that causes heat pattern/ syndrome, also called pathogenic heat”, which also includes **Fire** “as a pathogenic factor characterized by intense heat that is apt to injure fluid, consume qi, engender wind, inducing bleeding, and disturb the mental activities, also called pathogenic fire”

and

Wind: (external) “as a pathogenic factor characterized by its rapid movement, swift changes, and ascending and opening actions”, also called pathogenic wind or **(internal)** “the same as liver wind, wind in the interior due to abnormal movement of body’s yang qi”.

A review of the Descriptor definitions just provided shows that they are indeed very different to the Western medical definitions of the organs. They are more ‘poetic’ than anatomical and very many of the Descriptors have no correlation in Western medicine at all. This brief introduction to certain aspects of CM strongly shows the different approach to viewing the body that CM has when compared to Western medicine.

Each Descriptor is scored with a 0-5 Likert scale[101], where 0 represents the absence and 5 represents the maximum expression of problems in this Descriptor. This is an important innovation over the current approach of CM diagnosis practise of stating a primary, secondary, tertiary or even more patterns. The use of a linear scale also enables the utilisation of weighted versions of percentage and chance-removed agreement statistics such as Gwet’s AC2[50] to report agreement. Typically in contemporary CM a pattern is either recorded or not, with no gradation of severity easily determined. This is overcome with the strategy employed with the CMDD.

Descriptor scoring within the CMDD enables a Total Pathogenic Score (TPS) to be calculated. The TPS is the sum of scores from all Descriptors, and can

be used as a generalised wellness measure of a patient within CM terms. The TPS can be used to reliably track *changes* in the overall health of a patient. The TPS may have similar value in the CM profession to the SF-36, a validated Western medical wellness questionnaire[83-85]. As with the SF-36, score changes upon observed changes in a subject's health status recorded by the same practitioner should be more useful for the determination of health changes than the absolute score[84].

A feature of the CMDD is that any patient's zang fu diagnosis, consisting of either a combination of many patterns or simple single pattern diagnosis, can be mapped to the CMDD format.

6.1 Mapping CM Diagnoses to the CMDD

An example of a CM diagnosis mapped to the CMDD format will now be presented, and the diagnostic matching capacity of each system will be shown. In this hypothetical example, three raters gave diagnoses to a subject using the CM format, and these diagnoses were then mapped to the CMDD format. The Descriptor matches were compared in both formats. Matching in the CM format was scored according to the number of CMDD Descriptors used to form the matched CM patterns.

The principles used for Descriptor matching in CM versus CMDD formats are demonstrated in the table 6.1.1 and 6.1.2 examples.

Table 6.1.1 CM diagnoses

Rater 1	Rater 2	Rater 3	Matches
Lung Phlegm Cold		Lung Damp Heat	
Kidney Jing Xu		Kidney Yin Xu	
	Heart Phlegm		
Liver Qi Stasis	*Liver Qi Stasis*	*Liver Qi Stasis*	3
Qi and Xue Stasis	*Qi and Xue Stasis*		1
			4

The Descriptors shown in italics are matches; with the number of CM complete matches displayed in the final column. If three raters matched a score of three was given per descriptor, as A matched to B, A-C and B-C, while if two raters matched a single match was recorded, as in the above case Qi and Blood Stasis. Simple agreement in this example is calculated as four matches of a possible nine, or 0.44.

Table 6.1.2 presents the diagnoses from Table 6.1.1 mapped in the CMDD format. For example, Lung Phlegm Cold in CM maps to the descriptors Lung, Damp and Cold in the CMDD, similarly Lung Damp Heat maps into Lung, Damp and Heat. The hypothetical example presented in Table 6.1.1 and mapped to CMDD in Table 6.1.2 demonstrate most of the other rules applied for descriptor matching.

Table 6.1.2 CM mapped to CMDD - descriptors selected by raters in ***bold italics***

Rater 1	Rater 2	Rater 3	Matches
Liver	***Liver***	***Liver***	3
Kidney	Kidney	***Kidney***	1
Lung	Lung	***Lung***	1
Spleen	Spleen	Spleen	
Heart	***Heart***	Heart	
Qi Xu	Qi Xu	Qi Xu	
Yang Xu	Yang Xu	Yang Xu	
Yin Xu	Yin Xu	***Yin Xu***	1
Blood Xu	Blood Xu	Blood Xu	
Qi Stag	***Qi Stag***	***Qi Stag***	3
Xue Stag	***Xue Stag***	Xue Stag	1
Damp	***Damp***	***Damp***	3
Wind	Wind	Wind	
Heat	Heat	***Heat***	
Cold	Cold	Cold	
			13

Liver Qi Stasis consists of two Descriptors in the CMDD format, so when two raters select it, two matches occur rather than one in the CM. This is a distinct advantage over agreement in the CM format. The amalgamations used within the CMDD of Phlegm to Damp and Jing to Yin also facilitate further agreement. The increased number of matches, from 4 to 13 in tables 2 and 3 respectively, demonstrates the potential for increased agreement with the CMDD approach.

It should also be noted that if the CMDD format is used to define a patient's heath status, the deliberate non-selection of Descriptor by a practitioner is an indication of the status of the patient. Should two raters not select a particular

Descriptor then an agreement is deemed to occur. In the above example this would have provided another 18 matches, were it known that the non-selection was deliberate. Combined with the matches a total of 31 from a possible 45 are obtained, a simple agreement of 0.69.

Suppose that an individual is exceedingly healthy and has no health problems whatsoever. Practitioners examining such a subject should not select any Descriptors to define their health status. The non-selection of patterns is therefore as important as the inclusion of a Descriptor. For the above individual, high levels of agreement should be very high due to non-selection of factors by the raters. The recording of absence of disease factors is central to the CMDD approach and is a major point of difference from the conventional CM approach.

6.2 CMDD Scoring Example

Presented now will be a CMDD scoring example. The patterns chosen by rater one are mapped with scores allocated to each Descriptor in Table 6.2.1 and then presented in Figure 6.2.1. This example demonstrates some of the guidelines used when mapping CM diagnoses to the CMDD format.

Table 6.2.1 CM to CMDD mapping example

	Liver	Lung	Kidney	Damp	Cold	Xue Stasis	Qi Stasis	Yin Xu
Lung Phlegm Cold		3		3	2			
Kidney Jing Xu			2					3
Liver Qi Stasis	2						4	
Qi and Xue Stasis						2	3	

The number ascribed to each Descriptor indicates the level of disturbance for that descriptor and the non-selection of a Descriptor is also relevant. Whenever a Descriptor is included in more than one CM pattern, the highest value for repeated Descriptors is used. In the above example, Qi Stasis was part of two patterns with differing scores. A pattern can also be made up of descriptors with different values. Liver Qi Stasis shows an expression of greater Qi Stasis as indicated by a score of 4, while the Liver component is only scored as 2, representing its lesser manifestation. Descriptor scoring allows objective tracking of a patient's progress in individual Descriptors upon treatment; an invaluable added feature of the CMDD format.

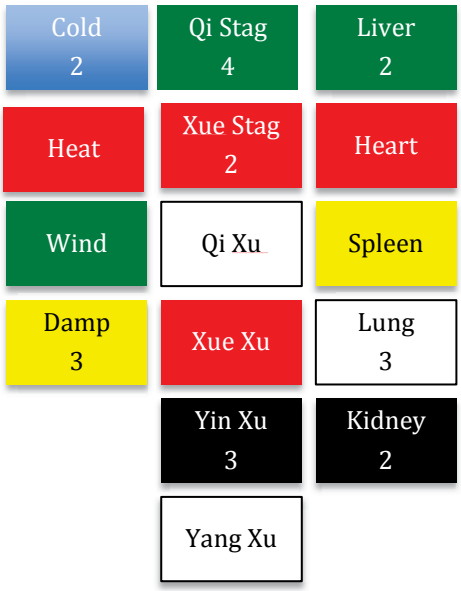

Figure 6.2.1 CMDD pattern mapping example

Additionally, Descriptor scoring within the CMDD enables a Total Pathogenic Score (TPS) to be calculated. The TPS is the sum of scores from all Descriptors, and can be used as a generalised wellness measure of a patient within CM terms. In the example shown in Figure 2, the TPS value is 21. The TPS can also be used to reliably track *changes* in the overall health of a patient, where a lower TPS indicates an improvement in health and vice-versa. Score changes in a subject's health status recorded by the same practitioner should be more useful for the determination of health changes than absolute Descriptor scores[84], due to possible practitioner scoring biases.

6.3 CMDD Agreement Calculation

Using the mapping example data presented in Table 6.2.1 and adding scores to the second and third practitioner's selections, an example of the method

used to calculate agreement will now be given, with the assumption that the zeros indicate that the Descriptor was deliberately omitted by the diagnosing raters, is presented in Table 6.3.1.

Table 6.3.1 Agreement Calculation Data

Descriptor	Rater 1	Rater 2	Rater 3
Liver	2	4	3
Kidney	3	0	4
Lung	2	0	3
Spleen	0	0	0
Heart	0	3	0
Qi Stag	4	4	2
Xue Stag	2	3	0
Yin Xu	3	0	2
Blood Xu	0	0	0
Qi Xu	0	0	0
Yang Xu	0	0	0
Cold	2	0	0
Heat	0	0	3
Wind	0	0	0
Damp	3	3	3
TPS	21	17	20

In the example linearly weighted simple agreement of 0.8 ±0.1 and linearly weighted Gwet's AC2 0.6 ±0.1 were found. The high standard error is due to having data for a single subject only. Higher levels of agreement are found here than the simple agreement of 0.44 found in the original CM diagnosis calculated in 6.1.

The CMDD was developed from the DSOM through a comparison to the UTS outpatient clinic data and an analysis of diagnoses recorded in the DSOM

data set made by practitioners trained in the contemporary Chinese Style of CM[71]. This causes the CMDD to be specific to the contemporary CM format.

If another data set of patient diagnoses were used sourced from another style of Traditional East Asian medicine, perhaps a different version of the CMDD would have resulted. There are different styles of CM; the most obvious example is Japanese meridian therapy, which would most likely need fewer Descriptors to adequately describe all diagnostic patterns due to the apparent simplicity of that system. It may be that the Descriptors of the DSOM were indeed appropriate for describing all patterns recorded in Korean style CM.

Examinations of large diagnostic databases from different styles of acupuncture would have to take place to determine if the Descriptors of the CMDD are appropriate to map all common diagnoses and experiments carried out to determine the inter-rater agreement between practitioners in each style. It could be that the CMDD would have to be modified for effective reporting of diagnoses in different CM styles. The testing of the CMDD format to ensure that it can be used with all the major styles of CM is another work in itself and beyond the present thesis.

Having developed the CMDD and shown that any zang-fu CM diagnosis in contemporary CM can be mapped onto the CMDD and presented an example of a method for calculating agreement, it remained to be shown whether agreement is improved with its use with real practitioner data, as appeared to be the case in the hypothetical example above.

Chapter 7 CMDD and CM Diagnostic Agreement Investigation

A second data collection was conducted which examined diagnostic agreement using the CMDD and that found with a contemporary CM diagnosis that was usually recorded in clinics. As with the first data collection to investigate diagnostic reliability using the DSOM format described in chapters 3-6, no treatment intervention was offered.

Two objectives were attempted in the second data collection: to determine the agreement usually obtained in as close to a normal clinical setting as could be possible, from an open population of subjects, using a format that approximated the contemporary CM diagnostic format, and to compare this agreement with that obtained with the same subjects using the CMDD. The second data collection also looked to validate the CMDD as a diagnostic format.

7.1 CMDD and CM Diagnostic Data Collection Details

7.1.1 Subjects

Volunteer subjects were recruited by word of mouth. In total, 35 participants were enrolled in the study (23 females, 12 males). There were no exclusions on the basis of age or health status. The mean age of the females was 50 and the males 56, with a range of 17-78 years.

Prior to commencing the study, ethical approval was obtained from the University of Technology, Sydney (UTS) Human Research Ethics Committee 2007-152[102].

Subjects were allocated to one of three appointment slots on either day. The order in which the subjects were interviewed was quasi-randomised by the order of arrival. This approach also assisted workflow. Prior to commencing data collection, each participant was asked to read an information sheet and sign a consent form.

7.1.2 Practitioners

On the first day four practitioners diagnosed nineteen subjects. On the second day six practitioners diagnosed sixteen subjects, leading to a total of 35 subjects. A total of 172 diagnostic assessments were therefore completed, 86 each in CM and CMDD formats.

While all efforts were made to recruit six practitioners for each day's data collection, therefore enabling the planned comparison of three practitioners utilising each format on each day, we were unsuccessful. The attempt to recruit the required number of practitioners continued up until the eve of the first day's data collection. To abort the data collection at this late stage and try to reschedule the attendance of thirty-five subjects and the practitioners who had already agreed to attend would have likely led to similar or even worse attendance problems, so it was decided to proceed and collect the data. This disparity in the number of diagnosing practitioners utilised on each day

unfortunately necessitated each day's data to be processed separately, with an associated loss of statistical power.

7.1.3 Location

Data was collected over two days at the UTS CM outpatient clinic. This clinic is closed to the public on the weekends, which made it a convenient and appropriate location to conduct the interviews.

7.2 The CM and CMDD forms used in the data collection

Two forms were used to collect the data from the practitioners, one using the CM and the other the CMDD format. When using the CM format, practitioners were instructed to choose a maximum of three Diagnostic Patterns (DP) from a list of 56 DPs and use their normal method for determining a DP. Details of the 56 DPs and the mechanism that was used to determine the DPs that were included in the list have been outlined in chapter 5, and the list of DPs used is presented in Appendix 5. Each DP was scored on a Likert scale[101] of 0-5 to record the severity of pattern presentation.

In a contemporary CM clinic, there is no attempt to restrict diagnoses to to a list of options. This means any restriction in the options available to a diagnosing practitioner is a movement away from the native 'laissez faire' environment that they usually work in. It was therefore decided to attempt to reduce the options available to practitioners in a minor way, to reflect the nature of this environment. As shown in the literature review in section 2.3.4, in all previous CM diagnostic reliability studies published, there was always a

strict limit to the diagnostic options available to the practitioners, with an average of less than three options offered, while examining dozens of diagnostic facets individually. These diagnostic investigations were put forward as a quazi-investigation into overall CM diagnostic reliability. This approach does not at all reflect the diagnostic landscape in practice in the 'average' CM clinic.

The approach we took, of limiting a practitioners choice to a finite list of precise terms, albeit a reasonably long list, was one that would hopefully not move too far away from the normal CM method of recording a diagnosis.

Not using an approach that embodied the provision of a precise list of available options, but one that exactly reflected CM practice, with a totally unrestricted diagnosis being recorded by the practitioners, would have created problems.

There are two extremes that may be taken in determining what indeed constituted agreement between unstructured diagnostic records. On one hand, agreement may have suffered due to slight semantic differences in pattern definitions, if a strict interpretation of what constituted agreement was applied. On the other hand, if a more liberal approach was taken and similar terms were seen as the only requirement for agreement, then there could be issues with the 'rules' used to determine which diagnoses were indeed the 'same'. It can even be argued that one should not even try to read the mind of the diagnosing practitioners and make a sweeping generalised decision that

we knew what they were trying to say, when they expressed the diagnosis in their own preferred way.

To completely avoid the dillemma presented, a precise list of terms is a logical solution. Some of the diagnostic options within the list of 56 DPs were similar anyway, which in part reflects the semantic issues faced.

When the practitioners were required to make a CMDD-formated diagnosis, the CMDD form was used. This form is presented as Appendix 6.

7.3 Agreement Calculation in Chinese Medicine and the Chinese Medicine Diagnostic Descriptor Formats

Two methods were utilised to report levels of consensus, linearly weighted percentage agreement, and chance-removed, linearly weighted agreement calculated by Gwet's AC2 statistic, a superior method to other comparable statistics as reported[38-40, 42, 47, 103]. The strength of this statistic for calculating agreement between multiple raters who have recorded scores on linear scales has been discussed in chapter 2.

Weighted statistics are relatively under utilised in inter-rater agreement, with only two citations observed in the literature[26, 104] compared to the greater number that used non-weighted[20-23, 25, 27, 28, 57]. Weighting of agreement can be one of three general approaches: quadratic, linear or radical. Quadratic is biased towards higher agreement with score proximity, while radical produces lesser agreement with score propinquity. Linear reports the intermediate

between these two approaches. The reasons for the selection of linear weighting to report agreement are outlined further in chapter 2.

A key difference between the calculations in the two formats is the inclusion of agreement in the CMDD format where a Descriptor is not scored. As the CMDD consists of the least Descriptors possible, and all could be selected, non-selection of a Descriptor becomes relevant. Agreement was calculated with this approach within the CMDD and the CM mapped to CMDD data. The mapped data showed similar trends using the same assumptions. The inclusion of non-selected patterns into an inter-rater agreement calculation is not feasible within the CM diagnostic format, as this diagnostic format comprises 56 patterns, and only a maximum of three patterns were instructed to be selected. Non-selection of a pattern in this format therefore cannot be utilised as a factor in agreement.

Chapter 8 Chinese Medicine Format Agreement

8.1 Initial Observations

The following observations were made regarding the data collected from the practitioners recorded with the CM format.

8.1.1 CM Patterns Chosen by Practitioners

Fourteen patterns were not used and a further five were only selected once. This means approximately one third of the DPs available for selection were either used once or not used at all. In contrast, the ten patterns most selected made up 59% of all selections.

8.1.2 Data Collected Compared with UTS Clinic Data

In a similar fashion to the comparison made between the data collected in the DSOM data collection and the UTS clinic data, previously outlined in section 5.2, data from the present study was compared with the records collected at the UTS clinic, to determine whether the patterns selected by the practitioners were representative of those generally seen in clinical settings in Australia.

Table 8.1.2.1 Comparison between the diagnoses selected at UTS outpatient clinic and the present study

	Present study				**UTS Student Clinic Data**			
Patterns	%	cumulative	selected	Rank	%	cumulative	selected	Rank
Liver Qi Stagnation	14%	14%	43%	1	13%	13%	19%	2
Spleen Qi Xu	10%	24%	33%	2	13%	26%	20%	1
Kidney Qi Xu	5%	29%	19%	3	6%	32%	9%	5
Qi & Blood Stagnation	5%	35%	18%	4	9%	41%	13%	3
Kidney Yang Xu	5%	39%	13%	5	1%	42%	2%	19
Kidney Yin Xu	4%	43%	12%	6	6%	47%	9%	6
Blood Xu	4%	48%	11%	7	2%	49%	2%	15
Wood (liver) invades Earth (spleen)	3%	51%	8%	8	2%	51%	3%	12
Kidney Jing Xu	3%	53%	8%	9	1%	52%	1%	29
Qi Stagnation in Gall Bladder Channel	3%	56%	8%	10	3%	54%	4%	9
Lung Qi Xu	2%	58%	7%	11	3%	57%	4%	8

Table 8.1.2.1 Comparison between the diagnoses selected at UTS outpatient clinic and the present study (continued)

Qi & Blood Stag in Gall Bladder Channel	2%	61%	7%	12	5%	62%	8%	7
Liver Yin Xu	2%	63%	7%	13	2%	63%	2%	14
Cold Damp in the Spleen	2%	65%	6%	14	1%	64%	1%	25
Heart Qi Xu	2%	67%	6%	15	1%	65%	1%	30
Liver Heat Rising	2%	69%	6%	16	2%	67%	2%	13
Qi Stagnation (localised trauma)	2%	71%	6%	17	0%	67%	1%	48
Spleen Yang Xu	2%	73%	6%	18	0%	67%	1%	43
Stomach Heat	2%	75%	6%	19	0%	68%	1%	49
Damp Heat in the Spleen	2%	76%	5%	20	1%	69%	2%	18
Heart Yin Xu	2%	78%	5%	21	1%	70%	1%	34
Heat in the Blood	2%	80%	5%	22	1%	70%	1%	31

Table 8.1.2.1 Comparison between the diagnoses selected at UTS outpatient clinic and the present study (continued)

Liver Wind (Internal - moving)	2%	82%	5%	23	1%	71%	1%	33
Qi & Blood stag in Colon Channel	1%	83%	5%	24	2%	72%	2%	16
Qi Stagnation in the Bladder Channel	1%	85%	5%	25	2%	74%	3%	10
Qi Stagnation in the Liver Channel	1%	86%	5%	26	1%	75%	1%	35
Damp Heat in the Bladder	1%	87%	4%	27	0%	75%	1%	37
Damp Heat in the Gall Bladder	1%	89%	4%	28	1%	76%	1%	23
Damp Heat in the Liver	1%	90%	4%	29	1%	77%	1%	28
Liver Blood Xu	1%	91%	4%	30	1%	78%	2%	21
Liver Yang Rising	1%	93%	4%	31	1%	79%	1%	24

Table 8.1.2.1 Comparison between the diagnoses selected at UTS outpatient clinic and the present study (continued)

Heart Fire blazes upwards)	1%	94%	2%	32	0%	80%	1%	41
Heat in the Heart	1%	95%	2%	33	0%	80%	1%	50
Heat in the Stomach	1%	95%	2%	34	0%	80%	1%	43
Liver Fire (blazes)	1%	96%	2%	35	0%	81%	1%	38
Phlegm Cold Obstructing the Lung	0%	97%	2%	36	0%	81%	0%	52
Cold Damp in the Large Intestine	0%	97%	1%	37	0%	81%	0%	56
Heart and Kidney not Communicating	0%	98%	1%	38	0%	82%	1%	46
Lung Yin Xu	0%	98%	1%	39	0%	82%	1%	36
Qi & Blood Stag Small Intestine Channel	0%	99%	1%	40	1%	84%	2%	17
Wind Cold Attacks the Lung	0%	99%	1%	41	1%	85%	2%	20

Table 8.1.2.1 Comparison between the diagnoses selected at UTS outpatient clinic and the present study (continued)

Bladder Qi Xu	0%	100%	0%	42	0%	85%	0%	54
Cold Stagnation in Colon Channel	0%	100%	0%	43	0%	85%	0%	55
Damp Heat in the Large Intestine	0%	100%	0%	44	0%	86%	1%	42
Heart Blood Xu	0%	100%	0%	45	0%	86%	1%	51
Phlegm Heat Disturbing the Heart	0%	100%	0%	46	1%	87%	1%	22
Phlegm Heat Obstructing the Lung	0%	100%	0%	47	0%	87%	0%	53
Qi & Blood Stag in the Bladder Channel	0%	100%	0%	48	7%	94%	10%	4
Qi & Blood Stag in the Spleen Channel	0%	100%	0%	49	0%	95%	1%	45
Qi & Blood Stag in the Stomach Channel	0%	100%	0%	50	1%	95%	1%	32

Table 8.1.2.1 Comparison between the diagnoses selected at UTS outpatient clinic and the present study (continued)

Qi Stagnation in Large Intestine Channel	0%	100%	0%	51	1%	96%	1%	26
Qi Stagnation in Small Intestine Channel	0%	100%	0%	52	1%	97%	1%	27
Qi Stagnation in the Stomach Channel	0%	100%	0%	53	0%	97%	1%	38
Spleen Blood xu	0%	100%	0%	54	0%	98%	1%	47
Stomach Qi xu	0%	100%	0%	55	0%	98%	1%	38
Wind Heat Attacks the Lung	0%	100%	0%	56	2%	100%	3%	11

Notable differences between the two data sets are the non-selection of *lung wind heat* and *wind cold* DPs in the present study. Since no treatment was offered as part of this study, subjects suffering such acute conditions were unlikely to have attended. Another observation was the low selection of *Bladder Qi* and *Blood Stagnation* in the present study. There is no explanation for this variation. Otherwise, there are great similarities between these two data sets.

8.2 Agreement calculation in the CM diagnostic format

Two approaches were used to determine agreement with the CM diagnostic format. In the first approach, termed Pattern Agreement, was deemed to have occurred where more than one rater recorded a score for a DP of 1 or greater for the same subject, irrespective of the score. This approach is the lowest possible threshold of agreement. The second approach, termed Weighted Agreement, compared the scores of each practitioner for each pattern matched as a linearly weighted percentage. The values shown below in Table 8.2.1 were used to calculate Weighted Agreement.

Table 8.2.1 Linear Weights Utilised

	1	**2**	**3**	**4**	**5**
1	1	0.8	0.6	0.4	0.2
2	0.8	1	0.8	0.6	0.4
3	0.6	0.8	1	0.8	0.6
4	0.4	0.6	0.8	1	0.8
5	0.2	0.4	0.6	0.8	1

As discussed in Chapter 2, the calculation of Simple Agreement always involves determining the obtained agreements and dividing this number by the number of possible agreements.

In the case of the diagnostic pairs in the 19 subjects who were diagnosed by pairs of practitioners on the first day, there were three potential agreements for each subject, or 57 possible agreements. On the second day's data collection, where three practitioners diagnosed the 16 subjects, nine agreements were possible in each subject, leading to a total of 144 potential agreements. Totaling both days potential agreements led to a total of 201 Possible Agreements.

The raters occasionally digressed from the guidelines of attributing three patterns per subject. On ten occasions only two patterns were ascribed. No DPs were ascribed at all by one practitioner to one subject. In three cases raters used four patterns and in one case five patterns were used. The digressions represent a movement away from the instruction of about seven percent.

Questions arise; should the Possible Agreements used to calculate agreement change due to the variation in the number of patterns ascribed? In the case of less patterns being used, it can be argued that the non-selection was due to the practitioner determining no other pattern existed, so there should be no reduction. In the case of extra patterns being utilised, even though there is an increased likelihood of agreement due to more selections,

unless two practitioners selected more than three patterns (and this was not the case in the present study), there is no change to the total of potential agreements, so there will be no increase in the potential pattern agreement possible.

There were 46 occasions where the raters chose the same pattern disregarding their scores, 12 on the first day and 34 on the second. This data was used to calculate Pattern Agreement. When this score data was processed to adjust agreement using the linear weights presented in Table 8.2.1, a score of 38.8 was derived, which determined Weighted Agreement.

8.3 Chance-removed Agreement calculation and the CM diagnostic format.

Generally chance-removed agreement is the preferred method for calculating agreement between raters. The data recorded in the CM format could not be processed by the normal algorithms used to produce these statistics, due to problems with the treatment of non-selected patterns. The certainty of chance-removed or simple agreement therefore could not be estimated either.

8.4 Agreement Results with the CM diagnostic format

The results are reported in Table 8.4.1 which shows that pattern agreement on both days was 23% and when weighted 19%.

Table 8.4.1 Chinese Medical format Agreements

Statistic	Day 1	Day 2	Both Days
Pattern Agreement	21%	24%	23%
Weighted Agreement	20%	19%	19%

Very poor agreement was found in both the Pattern and Weighted Agreement processes. As mentioned above, agreement by chance should be removed for true agreement to be calculated.

The inability to calculate recommended statistics such as Gwet's AC2 weighted[47, 50] agreement was a serious shortcoming of the CM diagnostic format. However, if such a process could theoretically be performed, the agreement reported must be lower than that obtained by simple percentage as reported. This is supported by the data presented in Figure 2.3.2.1, which supports the proposition that chance-removed agreement would be essentially non-existent in a theoretical AC2 CM inter-rater agreement calculation, where simple percentage is only 20% or so. In section 2.3.2 which investigated previously reported AC2 agreements, it was observed that the lower the raw percentage agreement, generally the greater the reduction in AC2 agreement compared to raw percentage, so logically, if chance-removed agreement could be calculated, it would be Poor according to Landis *etal*[53] at near or even below zero. This suggests that agreement using the CM method of recording a diagnosis is associated with no worthwhile agreement, a troubling result.

Treatment perforce depends on diagnosis. There is currently no other avenue to determine the 'correct' diagnosis than by reliable judgement capable of replication by another rater in CM. It is therefore crucial that is there is an acceptable consensus between raters as a representation of repeatable diagnoses. The finding of such *Poor* diagnostic agreement in open populations by CM practitioners presented in Table 8.3.1 bodes ill for the confident adoption of, or comparison between practitioners' treatment strategies.

The level of agreement reported in this study is an unacceptable foundation for any practice or investigation of effectiveness of treatments. As a consequence, the current thrust for determining results of treatment effectiveness from large-scale studies using data from many practitioners[1, 2, 6] may be premature, leading to wrong conclusions. Such studies need to wait until diagnostic agreement has been markedly improved.

In particular, this might be the reason for MacPherson[6] finding that *"There was little evidence that different characteristics of acupuncture or acupuncturists modified the effect of treatment on pain outcomes."* Since there is no single cause for back pain, practitioners diagnose and treat according to that diagnosis. If for instance, there were two possible diagnoses, and a group of practitioners incorrectly diagnosed half the time in each diagnosis, the treatment effectiveness would be blurred accordingly. Further, some of the practitioners, not being completely sure of the diagnosis,

may have treated for both DPs just in case, that is, used the 'shotgun' approach. MacPherson's observation that more needles per treatment were associated with greater effectiveness is in accord with this hypothesis.

Very interesting research using physiological measurements are being applied for determining the mechanisms of acupuncture[105-107]. There is a possibility that otherwise unexplainable variations in these measurements could be correlated with CM diagnostic variables, with the possibility of objective diagnostic tools eventuating. For any correlation to be successfully attempted, an acceptable level of diagnostic agreement is essential.

A number of recent studies[6, 8] have questioned the validity of CM theories. However, if consensus of CM pattern identification is totally lacking, the possibility exists that wrong conclusions concerning the presently taught theoretical basis of CM may be reached and the classical perspective rejected prematurely. This process can only be adequately and fairly evaluated once adequate diagnostic consensus has been attained. Indeed, whilst these theories may not be germane, at least they need to be fairly evaluated. It is clear therefore that there is a priority to improve the levels of agreement currently observed in CM practice.

There appeared however, to be a kind of consensus buried within the CM diagnostic data. While some patterns were hardly or never selected, others were selected quite frequently and within these selections, some patterns matched at greater rates than others. The selection and subsequent matching

of these patterns was buried within the plethora of diagnostic choices available, which thwarted mature chance-removed calculation of agreement and forced the ignoring of the implied agreement that non-selection of patterns could represent. To investigate the characteristics of the matches that did occur, table 8.3.2 was compiled that shows the number of matches and selections in the CM patterns used in the study.

Table 8.4.2 Patterns that attracted matches and the number of selections

Pattern	Matched	Selected
Liver Qi Stag	15	36
Spleen Qi Xu	7	27
Kidney Qi Xu	3	16
Lung Qi Xu	3	6
Heart Qi Xu	2	5
Kidney Yang Xu	2	11
Qi & Xu Stag in Gall Bladder	2	15
Qi & Xu Stag	2	6
Blood Xu	1	5
Cold Damp in Spleen	1	3
Heart Yin Xu	1	4
Heat in the Blood	1	4
Kidney Yin Xu	1	10
Liver Heat Rising	1	5
Liver Wind	1	4
Qi Stag	1	5
Qi Stag in Gall Bladder channel	1	7
Qi Stag in Liver channel	1	4
Total	46	173

The proportion of selections and matches as percentages of the total number of raters could not be clearly calculated due to the differing numbers of practitioners that diagnosed each subject. Selections and matches occurred in

two scenarios; when two raters agreed, one match resulted and when three raters agreed, three matches were generated.

Some patterns were selected and matched more frequently than others. Liver Qi stasis and Spleen Qi Xu were clearly the most frequently selected and had the highest number of matches. These two patterns are well known and seem to be very common patterns in CM practice. Practitioners made matches in these two patterns on 15 and 7 occasions respectively. These matches accounted for 48% of the total accord. Selections of these two patterns occurred in 36 and 27 occasions respectively, or 36% of the total. This implies that these two patterns, while selected roughly a third of the time, accounted for approximately half of the diagnostic agreement recorded.

Matches were statistically more likely with a greater number of selections, so the case of three matches from just six selections of Lung Qi Xu, or two matches from only five selections of Heart Qi Xu are more compelling than the matching of the most popular selections.

Some patterns were often selected but matched infrequently, as in the case of Kidney Yin Xu, which had only one match but ten selections.

8.5 Agreement with the CM diagnostic format Conclusion

Two methodologies for determining diagnostic agreement using the contemporary CM methods for recording a diagnosis have been developed. Chance-removed inter-rater agreement cannot be calculated with the CM

format with open populations with unrestricted pathology for each subject. Standard error of the agreement calculations also cannot be determined. The level of agreement between experienced CM practitioners diagnosing such open populations is very poor. Unless improvements to the CM diagnostic framework are developed and validated that improve diagnostic agreement to acceptable levels, the investigation of mechanism of action, the possible rejection of classical approaches, or the testing of treatment effectiveness may lead to unsound and premature conclusions.

In the next chapter, the same subjects that received diagnoses with the CM format had diagnoses recorded by other practitioners utilising the CMDD and inter-rater agreement between the practitioners is calculated and reported.

Chapter 9 Chinese Medicine Diagnostic Descriptor Agreement

The raw data of the practitioners using the CMDD, hereafter referred to as the 'CMDD data set', will now be presented and discussed, commencing with aspects of the data collected, culminating in the detailed analysis of the levels of agreement between the practitioners. The sequence and format will be similar to that used when presenting and discussing the DSOM data set in chapter 4. Comparisons between the results in the two data collections will then be made at each step.

It needs to be mentioned that none of the practitioners raised any concerns or had difficulties in using the CMDD format; indeed, there were positive comments as to the ease of use and value of this scheme.

9.1. Data Recorded by the Practitioners in the CMDD study

The frequency and percentage that the practitioners selected the scores available were determined and presented in table 9.1.1. This data was then compared and contrasted to the DSOM data set in 9.1.2.

Table 9.1.1 Scores selected in the CMDD data set

Scores	0	1	2	3	4	5	All
Selected	850	86	125	134	91	19	1290
Percent Selected	66%	7%	10%	10%	7%	1%	100%

Like in the DSOM data set, the practitioners again clearly and overwhelmingly chose 0 as their outcome for the subjects, with the scores from 1-4 all scoring at or just below 10%, with the highest score receiving only 1% of selection.

When this data is compared to the DSOM data set, the following differences between percentages of score selections are found.

Table 9.1.6 Differences in percentage selection in the DSOM and CMDD data sets

Score	0	1	2	3	4	5	total
CMDD	66%	7%	10%	10%	7%	1%	100%
DSOM	59%	8%	9%	12%	9%	3%	100%
Difference	7%	-1%	1%	-2%	-2%	-2%	

Zeros scores accounted for 66% of those allocated, a seven percent increase from the DSOM data set. The CMDD data set contained an even greater proportion of zero scores than the DSOM data set, which might indicate that the subjects were perceived by the practitioners as generally healthier in this study than in the DSOM study.

Next, the scores selected in each Descriptor are presented to determine each Descriptor's individual score frequency.

Table 9.1.3 Scores selections by all practitioners in each Descriptor

	Zero	1	2	3	4	5	Selected
Liver	31%	10%	14%	22%	16%	6%	69%
Kidney	35%	10%	21%	19%	13%	2%	65%
Qi Stag	47%	8%	15%	16%	12%	2%	53%
Spleen	49%	13%	8%	17%	13%	0%	51%
Qi Xu	60%	3%	14%	12%	10%	0%	40%
Damp	63%	8%	12%	12%	5%	1%	37%
Yin Xu	64%	6%	10%	10%	8%	1%	36%
Xue Xu	64%	6%	8%	9%	10%	2%	36%
Heat	71%	7%	8%	7%	6%	1%	29%
Blood Stag	78%	6%	5%	12%	0%	0%	22%
Cold	78%	6%	12%	2%	2%	0%	22%

Table 9.1.3 Scores selections by all practitioners in each Descriptor (continued)

Lung	81%	8%	5%	5%	1%	0%	19%
Yang Xu	87%	3%	3%	5%	1%	0%	13%
Wind	90%	1%	3%	2%	3%	0%	10%
Heart	91%	2%	5%	2%	0%	0%	9%
All Descriptors	66%	7%	10%	10%	7%	1%	34%

Liver was scored most frequently, as were Kidney, Qi Stagnation and Spleen. These four Descriptors were scored over 50% of occasions. The same top five selections; Liver, Kidney, Qi Stagnation, Spleen and Qi Xu occurred in the DSOM data set as well, although in different orders. These 'top five' Descriptors seem to be the 'meat and potatoes' for describing patient's general CM constitutional patterns, while the rest of the Descriptors are like spices that distinctly 'flavor the dish' so to speak.

At the other end of the spectrum, Heart was selected least of all, and the bottom four; Heart, Lung, Yang Xu and Wind were selected on less than 20% of occasions. After ignoring the differences of Dryness and Phlegm in the Descriptor DSOM list and Wind in the CMDD, four of the bottom five in each format was the same except in order. With these differences considered, each format had Heart, Lung, Yang Xu and Cold in their least scored five. One minor difference was the CMDD data set had Blood Stagnation and the DSOM data set had Blood Xu in the least chosen category.

9.2 CMDD and UTS Student Clinic Data Compared

In a similar fashion to the process presented in section 5.2, the data collected at the UTS outpatient clinic was compared to the CMDD data to determine if

the data collected was representative of the UTS clinic population. This was achieved by ranking the two sets of Descriptors.

Table 9.2.1 Descriptor Rankings with the CMDD data and the UTS outpatient clinic

Descriptor	CMDD	UTS Clinic	Rank Difference
Liver	1	2	1
Kidney	2	4	2
Qi Stag	3	1	2
Spleen	4	6	2
Qi Xu	5	5	0
Damp	6	11	5
Yin Xu	7	9	2
Blood Xu	7	12	5
Heat	9	7	2
Blood Stag	10	3	7
Cold	10	13	3
Lung	12	8	4
Yang Xu	13	14	1
Wind	14		
Heart	15	10	5

The first five most frequently selected Descriptors were within two ranking points of each other. Notable differences occurred in blood stagnation, which was the third highest Descriptor at the UTS clinic, while only the tenth within the CMDD trial. This is understandable as patients were attending the UTS clinic for treatment, while the subjects attending the CMDD diagnostic trial were not. Blood stasis is associated with chronic pain[32], a common condition for which patients attend acupuncture clinics. Other descriptors with substantial differences were damp, blood deficiency and heart, each having a five-point difference in ranking. The correlation co-efficient calculated between these two Descriptor rankings was reasonable at 0.63.

9.3 CMDD Agreement Results

As with the DSOM data set, two methods were utilised to calculate and evaluate levels of consensus; linearly weighted percentage agreement, and chance-removed, linearly weighted agreement calculated by Gwet's AC2 statistic, a superior method to other comparable statistics as reported by the author and others[38-40, 42, 47, 103]. The results will be discussed using Landis' scale[53] for Kappa agreement.

As previously discussed in Chapter 4 when reporting agreement with the DSOM data, standard error calculated by the AC2 provides important additional information as to how reliable the agreement result is and gives the confidence of the result. A high standard error may mean that the sample is too small, or that agreement is tentative.

As previously discussed when the DSOM was introduced in Chapter 2 and again when the CMDD was introduced in chapter 6, a key difference between the calculation of CM and CMDD agreement is the inclusion of agreement in the CMDD format where a Descriptor is not scored. As the CMDD consists of the least Descriptors possible, and all could be potentially selected, non-selection of a Descriptor becomes as important as its selection. Agreement was therefore calculated with non-selection accord included within the CMDD. The mapped data showed similar trends using the same assumptions. This approach of pattern non-selection inferring agreement is not feasible within the CM diagnostic format, as this diagnostic format comprised 56 patterns,

and only three pattern selections were prescribed. Non-selection of a pattern therefore cannot be a factor in agreement within this format.

Linearly Weighted Simple and AC2 Agreements together with standard errors where available in the Chinese medical and CMDD formats are reported in table 9.3.1.

Table 9.3.1 Linearly weighted agreements calculated in the CM and CMDD formats

	Simple Agreement	Standard Error	AC2	Standard Error
CM day one	20	N/A	N/A	N/A
CM day two	19	N/A	N/A	N/A
CMDD day one	0.80	0.01	0.67	0.03
CMDD day two	0.79	0.01	0.67	0.03

Agreement established with linear weighted percentage and AC2 reported agreement, is immensely superior when the CMDD approach is used rather than the CM process. Using Landis and Koch's[53] accepted method of interpreting chance-removed agreement reported by Kappa statistics, the CMDD format demonstrate levels of AC2 agreement classified as Substantial, while the CM system would be classified as Slight, or if chance agreement could be calculated, maybe even Poor.

The chance-removed agreement of both days was slightly higher than that obtained in the DSOM data set of 0.60 ±0.02. The root mean square of the standard errors of 0.04 was calculated to determine the certainty of the difference between the inter-rater agreements calculated in each data set.

The certain difference between these two agreement calculations is reduced after consideration of this figure to a minor 0.03. The Substantial and repeated agreement calculated does however confirm that the superior underlying approach of the DSOM and CMDD formats for recording a diagnosis promote agreement between raters rather than the contemporary CM diagnostic format.

9.4 CM diagnoses mapped to the CMDD Format

The patterns chosen by the practitioners using the CM format were next mapped to the CMDD format using the rules previously outlined in the mapping example in section 6.1 of this thesis and linearly weighted percentage and AC2 inter-rater agreement were again calculated. The results are reported in 9.4.1.

Table 9.4.1 AC2 inter-rater agreement of CM diagnoses mapped to CMDD

	Simple Agreement	Standard Error	AC2	Standard Error
CM to CMDD day one	0.78	0.015	0.65	0.03
CM to CMDD day two	0.83	0.015	0.73	0.03

CM mapped to CMDD results demonstrate the level of agreement possible if the raters had recorded their diagnoses using the CMDD instead of the CM format and confirm the CMDD as a format that allows the true intentions of raters to be compared. Agreement in the CM to CMDD mapped data is effectively the same as if the CMDD were used to record these data, after square root of the sum of squares of standard errors, calculated to be 0.04 are considered.

The agreement calculated in the CMDD format mapped from the CM diagnoses suggests that CMDD seems to allow the true intention of the rater to be expressed in a format that allows appropriate inter-rate agreement to be calculated.

9.5 CMDD Agreement in the Three Wellness Groups

As with the DSOM data collection reported in 4.4, each subject's TPS was used for allocation into one of three approximately equal groups of subjects in both data sets. Linearly weighted AC2 chance-removed agreement was calculated for each subject by comparing the raters' scores of each of the Descriptors.

Table 9.5.1 Agreement in wellness groups CMDD data collection day one

Groups	Simple Agreement	Standard Error	AC2	Standard Error	Average TPS
Most Well	0.89	0.02	0.83	0.03	6
Intermediate	0.75	0.03	0.58	0.05	13
Least Well	0.78	0.03	0.59	0.06	18
All Groups	0.80	0.01	0.67	0.03	13

Table 9.5.2 Agreement in wellness groups CMDD data collection day two

Groups	Simple Agreement	Standard Error	AC2	Standard Error	Average TPS
Most Well	0.81	0.02	0.71	0.05	8
Intermediate	0.83	0.02	0.71	0.04	13
Least Well	0.74	0.03	0.57	0.06	15
All Groups	0.79	0.01	0.67	0.03	13

A similar pattern to the DSOM data collection was observed, with declining agreement in the Least Well group when compared to the Most Well.

9.6 Agreement in the CMDD's Individual Descriptors

As in section 4.6 of the DSOM data collection, simple agreement and AC2, average TPS and percentage of times chosen is now presented. The data is sorted by the AC2 agreement from lowest to highest values, to highlight the Descriptors that scored the lowest AC2 values and then presented in tables 9.6.1a and b.

Table 9.6.1a Individual Descriptor Agreement expressed in linearly weighted percentage and AC2, and the Average TPS and % of occasions the Descriptor was scored one or greater in day 1 data

Descriptor	Simple Agreement	Standard Error	AC2	Standard Error	Av TPS	% Zero choice
Qi Xu	0.60	0.08	0.19	0.18	1.58	47%
Liver	0.66	0.07	0.24	0.15	2.26	29%
Spleen	0.63	0.08	0.26	0.18	1.71	50%
Qi Stag	0.72	0.07	0.41	0.15	2.08	37%
Kidney	0.75	0.06	0.45	0.15	1.84	37%
Damp	0.73	0.06	0.56	0.14	0.89	68%
Heat	0.75	0.06	0.59	0.12	0.89	66%
Yin Xu	0.79	0.06	0.69	0.12	0.74	74%
Blood Stag	0.79	0.08	0.70	0.13	0.84	76%
Lung	0.79	0.07	0.72	0.12	0.63	82%
Yang Xu	0.88	0.06	0.83	0.09	0.82	76%
Cold	0.87	0.06	0.85	0.08	0.32	89%
Wind	0.92	0.06	0.91	0.07	0.21	95%
Heart	0.93	0.04	0.92	0.05	0.18	92%
Xue Xu	0.94	0.04	0.92	0.05	0.16	95%
Averages	0.78	0.06	0.62	0.12	1.01	68%

Table 9.6.1b Individual Descriptor Agreement expressed in linearly weighted percentage and AC2, and the Average TPS and % of occasions the Descriptor was scored one or greater in day 2 data

Descriptor	Simple Agreement	Standard Error	AC2	Standard Error	Av TPS	% Zero choice
Qi Stag	0.73	0.05	0.39	0.11	1.96	33%
Liver	0.76	0.04	0.45	0.09	1.85	33%
Kidney	0.76	0.05	0.56	0.12	1.1	56%
Blood Stag	0.73	0.06	0.59	0.13	0.9	71%
Qi Xu	0.81	0.04	0.65	0.1	1.02	56%
Spleen	0.81	0.05	0.65	0.13	1.21	56%
Yin Xu	0.82	0.06	0.69	0.12	1.04	67%
heat	0.83	0.05	0.7	0.11	1	63%
Heart	0.83	0.05	0.76	0.09	0.56	79%
Xue Xu	0.85	0.05	0.79	0.08	0.6	77%
Yang Xu	0.86	0.06	0.83	0.08	0.35	90%
Damp	0.9	0.05	0.88	0.07	0.38	85%
Lung	0.91	0.04	0.89	0.06	0.31	88%
Wind	0.93	0.04	0.92	0.05	0.23	92%
Cold	0.94	0.04	0.94	0.05	0.19	94%
Averages	0.83	0.05	0.71	0.09	0.85	69%

There were differences in the frequency of selection and levels of agreement in the data collected on each day. On day one, the same five Descriptors; Qi Xu, Qi Stagnation, Liver, Spleen and Kidney that were selected most frequently and performed worst as reported by AC2 statistic in the DSOM study presented in chapter 4, were again most selected and poorest inter-rater agreement performers. It seems that an average TPS of 1.5 or greater always leads to AC2 agreement of less than 0.50.

On day two, this correlation was repeated, with the two AC2 agreements below 0.50 occurring where the average TPS was again greater than 1.5.

At the other extreme, Almost Perfect agreement, as classified by Landis and Koch[53] as chance-removed agreement of 0.80 or greater occurred in almost every instance where the average TPS was less than 0.40. There were nine such cases and only one where average TPS was greater than 0.40.

The observations made in the DSOM data set regarding TPS and AC2 inter-rater agreement have been repeated. It seems that when a Descriptor is scored on average 1.5 or above, AC2 agreement undergoes reduced values.

9.7 CMDD Descriptor Agreement and TPS

Continuing with the same data analysis sequence as performed on the DSOM data in chapter 4.7; the subjects in each Descriptor were again sorted into three Wellness Groups according to average TPS. It was necessary to look at the data collected on each day separately due to the different numbers of practitioners used on each day. Due to this reason, as well as the higher numbers of zero scores recorded, the numbers for analysis in each day's data was less than optimal. Many Descriptors had insufficient scores in the Least Well and even the Intermediate category had insufficient scores in some Descriptors.

Even if the two days could be combined somehow, there would appear to be many Descriptors that would have not had more than three subjects with

average TPS scores that satisfied many of the Least Well and a few of the Intermediate scoring criteria, which was the minimum data required for an agreement calculation in any Descriptor Wellness cohort. Linearly weighted simple and AC2 agreement was calculated and was presented for both days in tables 9.7.1a and b and 9.7.2a and b.

Table 9.7.1a Intra-descriptor wellness groups Simple agreement day one

Groups	Least Well		Intermediate		Most Well	
Statistic	Simple Agreement	Std Error	Simple Agreement	Std Error	Simple Agreement	Std Error
Blood Stag	NA		0.20	0.07	0.97	0.03
Cold	NA		NA		0.98	0.02
Damp	NA		0.4	0.00	0.9	0.05
Heart	NA		NA		0.99	0.01
Heat	NA		0.40	0.00	0.86	0.05
Kidney	0.90	0.06	0.51	0.11	0.84	0.11
Liver	0.83	0.05	0.44	0.09	0.93	0.05
Lung	NA		0.33	0.06	1.00	0.00
Qi Stag	0.91	0.06	0.33	0.09	0.87	0.06
Qi Xu	0.87	0.11	0.17	0.05	0.84	0.05
Spleen	0.75	0.08	0.29	0.09	0.88	0.06
Wind	NA		NA		0.97	0.03
Xue Xu	NA		NA		1.00	0.00
Yang Xu	0.93	0.05	0.33	0.05	1.00	0.00
Yin Xu	NA		0.30	0.21	0.94	0.04
Averages	0.87	0.07	0.34	0.07	0.93	0.04

Table 9.7.1b Intra-descriptor wellness groups Simple agreement day two

Groups	Least Well		Intermediate		Most Well	
Statistic	Simple Agreement	Std Error	Simple Agreement	Std Error	Simple Agreement	Std Error
Blood Stag	NA		0.49	0.03	0.93	0.03
Cold	NA		NA		1.00	0.00
Damp	NA		NA		0.94	0.03
Heart	NA		0.60	0.05	0.90	0.05
Heat	NA		0.67	0.06	0.98	0.02
Kidney	NA		0.64	0.06	0.85	0.06
Liver	0.87	0.05	0.64	0.05	0.84	0.05
Lung	NA		NA		0.95	0.05
Qi Stag	0.84	0.03	0.56	0.07	0.84	0.05
Qi Xu	NA		0.76	0.06	0.84	0.06
Spleen	NA		0.67	0.06	0.95	0.05
Wind	NA		NA		0.96	0.02
Xue Xu	NA		NA		0.9	0.04
Yang Xu	NA		0.47	0.00	0.95	0.04
Yin Xu	NA		0.60	0.09	0.96	0.04
Averages	0.85	0.04	0.61	0.06	0.92	0.04

Table 9.7.2a Intra-descriptor wellness groups AC2 agreement day one

Groups	Least Well		Intermediate		Most Well	
Statistic	AC2	Std Error	AC2	Std Error	AC2	Std Error
Blood Stag	NA		-0.62	0.16	0.97	0.03
Cold	NA		NA		0.97	0.03
Damp	NA		0.4	0	0.88	0.07
Heart	NA		NA		0.99	0.01
Heat	NA		0.02	0	0.82	0.08
Kidney	0.84	0.11	-0.15	0.24	0.9	0.06
Liver	0.68	0.11	-0.25	0.16	0.92	0.08
Lung	NA		-0.4	0.01	1	0
Qi Stag	0.87	0.09	-0.4	0.02	0.81	0.11
Qi Xu	0.83	0.17	-0.64	0.13	0.74	0.1
Spleen	0.49	0.23	-0.56	0.13	0.84	0.1

Table 9.7.2a Intra-descriptor wellness groups AC2 agreement day one (continued)

Wind	NA		NA		0.97	0.04
Xue Xu	NA		NA		1	0
Yang Xu	0.89	0.09	-0.26	0.19	1	0
Yin Xu	NA		-0.56	0.47	0.93	0.05
Averages	0.77	0.14	-0.34	0.14	0.92	0.05

Table 9.7.2b Intra-descriptor wellness groups AC2 agreement day two

Groups	Least Well		Intermediate		Most Well	
Statistic	AC2	Std Error	AC2	Std Error	AC2	Std Error
Blood Stag	NA		-0.07	0.07	0.91	0.05
Cold	NA		NA		1	0
Damp	NA		NA		0.94	0.04
Heart	NA		0.15	0.09	0.89	0.07
Heat	NA		0.32	0.14	0.98	0.02
Kidney	NA		0.21	0.17	0.81	0.1
Liver	0.77	0.12	0.23	0.13	0.75	0.11
Lung	NA		NA		0.95	0.04
Qi Stag	0.72		-0.05	0.19	0.75	0.1
Qi Xu	NA		0.5	0.17	0.79	0.09
Spleen	NA		0.28	0.25	0.95	0.06
Wind	NA		NA		0.96	0.03
Xue Xu	NA		NA		0.88	0.06
Yang Xu	NA		0.19	0	0.94	0.04
Yin Xu	NA		0.04	0.21	0.95	0.05
Averages	0.75	0.09	0.18	0.14	0.9	0.06

Similar inter-rater agreement patterns to those observed with the DSOM data set occurred in the CMDD data set in the intra-descriptor wellness groups. The Least Well and Most Well groups obtained Substantial and Almost Perfect agreements, while the Intermediate group had Poor inter-rater agreement as reported by chance-removed linearly weighted AC2 statistics.

The CMDD actually performed a little better than the DSOM data set in the Least Well Descriptor and somewhat worse than the DSOM in the Intermediate groups but, as there was less data available in the CMDD data set, especially in the Least Well, these observations are not strongly supported.

There were more cases of Descriptors in the CMDD than in the DSOM data set that could not have agreement calculated in wellness groups. In total, there were nine missing Least Well Descriptor calculations on day one and thirteen on day two. Four Intermediate Descriptor calculations could not be performed due to insufficient subjects meeting inclusion criteria on both days. The total number of Descriptor calculations that were unable to be performed was therefore thirty a total number of Descriptor wellness combinations on both days of ninety. In contrast, there were eight missing Descriptor wellness calculations in the DSOM data set from the potential of forty-eight, all of which occurred in the Least Well category. This was a doubling of the proportion of missing Descriptor wellness combinations that were unable to be examined in the CMDD compared to the DSOM data set.

The reasons for the greater number of missing calculations in the CMDD data sets were as follows: firstly, there was the slightly smaller data set to begin with, 35 subjects compared to 42 with the DSOM. Secondly, the necessity to split the data set into two groups prior to examination led to much smaller samples of only 19 and 16.

9.8 Discussion of the results of the CMDD and CM data collections

Substantial inter-rater agreement of 0.67 ±0.03 was obtained in this experiment where the CMDD format was used to record diagnoses on both days. The agreement calculated corroborates the Substantial inter-rater agreement result of the DSOM data set.

The diagnoses recorded in the CM format were mapped to the CMDD format and 0.65 ±0.03 and 0.73 ±0.03 chance-removed agreements were calculated on each day, compared to the simple agreement of 19% calculated in the CM-formatted diagnoses with the same subjects. The mapping result confirms the capacity of the CMDD format to provide meaningful agreement calculations. These were the same raters' diagnostic choices that were used to reach such low agreement using the CM format.

The subjects in the CMDD dataset were sorted into three Wellness groups according to the TPS and inter-rater agreement between the practitioners was calculated for each group on each day. Almost Perfect 0.83 ±0.03 and Substantial chance-removed agreement of 0.71 ±0.02 was calculated in the Most Well Groups. Substantial 0.71 ±0.03 and Moderate agreements of 0.58 ±0.05 were estimated in the Intermediate Groups. Moderate 0.59 ±0.06 and 0.57 ±0.06 agreements were calculated in the Least Well Groups. Again similar to the DSOM data, the decline in inter-rater agreement in the Intermediate and especially the Least Well groups was noted.

Similar to the DSOM data set, when agreement was calculated in the individual Descriptors, where the average TPS was greater than 1.5 Descriptors had Slight to Fair inter-rater agreement with an average of 0.34 ±0.14. The pattern observed was that the higher the TPS, the lower the chance-removed agreement.

As with the DSOM data set, the raters' scores of individual Descriptors within the CMDD data set were sorted into intra-descriptor groups based upon the average TDS of each Descriptor. In the Most Well intra-descriptor groups, defined as with average TDS scores of one or less in each Descriptor, an average Almost Perfect chance-removed agreement of 0.92 ±0.05 and 0.90 ±0.06 on each day was calculated. In the Least Well intra-descriptor groups, defined as with average TDS of three or greater in each Descriptor an average Moderate agreements of 0.77 ± 0.14 and 0.75 ±0.09 were calculated.

Inter–rater agreement in the Intermediate groups, defined as average individual Descriptor TDS above one and below three, was Poor, with an average of -0.34 ±0.14 across all Descriptors. Based upon the observations made in the wellness groups when all Descriptors were included, one would expect that the agreement of the Intermediate intra-descriptor sub-groups to be between that calculated in the Most Well and Least Well.

This outcome, in the same pattern, but even more pronounced than that found in the DSOM data set is again counter-intuitive when the Wellness Group

results are considered. This confirms a new understanding that agreement is lowest where raters score descriptors in an intermediate fashion and higher when descriptors are scored more heavily or lightly. Agreement in the intra-descriptor wellness groups provides a new insight into what needs to be targeted to improve agreement.

The CMDD format appears to have many advantages over the incumbent CM structure used to describe the details as well as the totality of a subject's health. The increased agreement found in the validation study, supported with the agreement observations made in the DSOM data set combine to form a persuasive argument for its adoption to adequately describe the constitutional condition of a subject.

The increased agreement will be useful in research validating treatment, investigating mechanisms of CM action, as well as use in contemporary clinical settings. Other researchers need to validate the CMDD. Validated questionnaires for some of the diagnostic factors that could be applied to the Descriptors have been published[59, 77-80], a completion of questionnaires for all Descriptors would dovetail nicely with the CMDD approach, potentially further improving agreement.

The CMDD would provide a simple definition of the CM pattern underlying each disease, and therefore facilitate diagnostic agreement. Benchmarks of current levels of diagnostic reliability need to be established. The diagnostic reliability of practitioners used in research should be reported as part of

upgraded Standards for Reporting Interventions in Clinical Trials[29] (STRICTA) or the Consolidated Standards of Reporting Trials[30] (CONSORT) research guidelines. Strategies should be devised and tested to attempt to further push up the levels of diagnostic agreement to ever-greater levels.

The empirical evidence of thousands of years of CM diagnosis and treatment practice could be properly evaluated, and perhaps reinterpreted. It may be that some or even all of the observations previously recorded are found to be without significance, but to discard the enormous repository of clinical observations we have been bequeathed by the CM profession, without proper evaluation, would indeed be irresponsible.

The CMDD shows promise as a validated diagnostic system and allows evaluation of the genuine inter-rater consensus; while the current overly semantic, strict diagnostic definitions currently used[70, 71] do not. This outcome needs to be further verified, as a small data set was used for this initial investigation.

The CMDD also appears to be a superior system for recording diagnoses, and it should be adopted across the CM profession. Improved pattern agreement and the capacity of tracking of a patient's health after treatment would follow. Research into the effectiveness of treatment strategies or mechanisms of CM action should use the CMDD as the diagnostic reporting format, and report the estimated diagnostic certainty of the subjects CM

condition by pre-testing and reporting the diagnostic reliability of the practitioners involved.

Chapter 10 Normalisation of data: attempt at partial removal of practitioner bias

In chapter 9, the Chinese Medicine Diagnostic Descriptor (CMDD) format was developed and validated. It was shown to be an excellent diagnostic format, in some ways possibly superior to the standard CM arrangement examined in chapter 8. A Substantial[53] chance-removed linearly-weighted agreement of 0.67 ±0.03 was obtained on both days data using Gwet's AC2[47] statistic. This result is in agreement with the result of 0.60 ±0.02 obtained with the DSOM format on a similar subject population presented in Chapter 4. In Chapter 8, simple linearly-weighted agreement of only 0.19 was found between CM practitioners using the contemporary CM diagnostic format. Whilst the results using either the CMDD or DSOM formats are superior to the results obtained with conventional CM diagnostic tools, as shown in chapters 4 and 9, the large differences in TPS of each subject allocated by different practitioners using CMDD or DSOM formats indicate that agreement could be further improvement.

In sections 4.7 and 9.7 more detailed investigations were performed concerning diagnostic agreement within each Descriptor in the DSOM and CMDD data sets. Three intra-Descriptor wellness groups were created within each Descriptor according to TDS of the subjects.

The agreement between raters of the intra-Descriptor sub-groups of the subjects classed as Most Well was in most cases ranked as Almost

Perfect[53]. The Least Well intra-Descriptor groups, where sufficient data were available that met the TDS inclusion criteria, achieved Moderate[53] agreement. The agreement Intermediately Well intra-Descriptors however, generally rated as Slight[53] .

The Almost Perfect and the Moderate levels of inter-rater diagnostic agreement obtained in the Most Well and Least Well groups of subjects respectively are positive outcomes that should not be overlooked. It indicates that agreement when patients are well is robust when either the DSOM or CMDD formats are used. This outcome supports the view that the CM profession is able to correctly agree where high levels of wellness *or* serious chronic disease are present and is clearly facilitated by the use of an appropriate diagnostic format such as the DSOM or CMDD.

The low level of intra-Descriptor chance-removed agreement between raters of the Intermediate groups is a cause for unease. These are the patients that practitioners need to treat effectively to prevent escalation to chronic diseases. Indeed, how can treatments of subjects with moderately poor levels of health be confidently applied? Therefore how can treatment outcomes arising from these uncertain diagnoses be correctly interpreted?

Work must be done to improve inter-rater agreement, for the unwell and especially for the moderately unwell subjects. While the employment of the CMDD or DSOM styles of format is an excellent start, improvements in agreement must still be pursued. If diagnostic agreement can be improved,

the results of interventions, or any examination of the mechanism of action of CM based upon this foundation can be tested with confidence with subjects.

As discussed in section 2.5, Ward *et al*[66] found that the use of an overly complex and indistinct diagnostic terminology to describe illness, led to rater bias in its many forms. This was the second most significant cause for disagreement in psychiatry: accounting for 17% of practitioner disagreement in their study. If improvements in diagnostic agreement of this magnitude were to be achieved through some process of bias removal in the present data, a significant gain in inter-rater agreement would result. In this chapter a number of attempts are made to develop processes to explore whether a consistent bias exists in a rater's scores and whether it is possible to remove it once it is detected.

10.1 Utilisation of the DSOM data instead of the CMDD data

The DSOM format data previously presented in Chapter 4 was used in an investigation into bias and agreement. This was not the preferred choice of data, but the use of a different number of raters on each day when the CMDD format was used, means that CMDD data would have to be processed separately for each day, leading to two smaller samples of 19 and 16 each.

When the DSOM format was employed, five raters were present at all times and a large amalgamated DSOM data set was constructed so that this data

set was used in the study of bias. More reliable statistical results could therefore be obtained.

10.2 Bias Measurement Theory

In a manner similar to the approach used when dealing with errors in the physical sciences, the score allocated by any rater can be considered as consisting of a combination of the "true score" and a measurement error. A true score theory is explained by Trochim[108] as "every measurement is an additive composite of two components: **true ability** (or the true score) of the respondent on that measure; and **error".**

This error can be further divided into two components, random error and systematic error. Random errors are products of chance, and therefore should not affect the mean, but the variability of a parameter is directly dependent on that error. If the error is truly random, then the parameter will be equally distributed about the mean. Systematic error on the other hand, will tend to push a measurement or score consistently in one direction, and therefore in the cases where raters are involved is bias. Examples of causes of bias in are changed conditions of data measurement changes such as noise or temperature in the data collection environment, or possibly more likely the subjectivity of raters.

Subjectivity of raters can take many forms and is defined here as an expression of the personal qualities of the rater. These include their personal

views, experience and background. If there were a systematic bias component within any practitioner's scores, then its removal should in principle improve agreement between raters. The removal of systematic errors from the scores allocated by raters is the goal of this investigation

This formed the following hypothesis; if a component of raters' bias was consistent, then its identification and removal would improve agreement.

10.3 Score Bias and Descriptor Bias Approaches to Normalisation

The first attempt, called Score Normalisation is based on the assumption that a rater is *always biased in the same way*; that is he/she rates *all Descriptors* either high or low. The second, designated Descriptor Normalisation is based on the assumption that a rater tends to be *only biased in rating certain Descriptors in the same way:* that is he/she rates *certain Descriptors* either high or low. Both types of normalisations required the DSOM data recorded on each day to be processed separately as different combinations of practitioners were used on each occasion.

Normalisations were only applied to the non-zero scores, as a zero score was considered an absolute choice indicating that the condition described by the particular descriptor was definitely not present in the subject. The percentage of zero scores in the data was 59% on day one, 67% on day two and 47% on day three and 59% overall. Consequently, the number of scores that would be changed by normalisation is less than half of the available data. Also scores

for individual descriptors were not allowed to be reduced below zero or increased beyond five to retain the original score choice range after any normalisation process was implemented. The significant number of zero scores in the data reduced the effectiveness of the normalisation effect in all the normalisation processes carried out in the present work.

The number of zeros present in the data was greatest in the subjects that were classified as Most Well with 73% of all diagnostic choices made being zero. The Intermediate and Least Well groups on the other hand, had 60% and 47% respectively. This logically means that the normalising process would leave Most Well group relatively unchanged, whilst the Intermediate and Least well groups would be progressively more affected.

Scores were also never allowed to be reduced below zero or increased beyond five to retain the original score choice range after any normalisation process was implemented.

Score Normalisation was the simplest approach to the removal of bias attempted and was designed to overcome each practitioner's potential propensity to score either aggressively or conservatively in a generalised way.

The Descriptor Bias approach tested whether bias may differ in each component of the diagnosis recorded by the raters. This was a more detailed investigation, which caused the scores of each Descriptor to be normalised individually and was investigated after the Score Bias approach.

Two approaches to the Score and Descriptor Bias normalisations were made, leading to four normalisation attempts in total. The first was called Trip Factor Normalisation and the second was called Score Factor Normalisation. Both slightly different approaches will be described fully in the appropriate sections and both were designed to adjust the scores of each practitioner so that averages based on the adjusted scores were approximately the same for all practitioners in slightly different ways.

The Normalisation implementation processes will be presented in the following pattern in sections 10.4 to 10.8. After introducing and describing a particular approach, the equation for deriving the normalisation factors applied to the raw scores will be presented in section 10.x.1. Here the symbol x represents the section number within this chapter so that x = 3 in the present section Following this in 10.x.2 will be the calculation of the normalising factor values with the relevant equation and raw data. This is followed by section 10.x.3 in which a summary of the Normalised data where score and percentage differences between the raw and normalised scores is presented. In the case of the Descriptor normalisations, the tables used to calculate the normalising factors 10.x2 and the score changes after normalisation implementation were quite large, so were included in the appendices. Final sections 10.x.4 in each Normalisation instance present the overall agreement and agreements in each Wellness group after Normalisation. Agreement calculated after Normalisation and the raw data will be compared to determine

if significant changes were observed after normalisation and the results discussed.

10.4 Score Bias Normalisation by "Trip factor"

Score Normalisation by the Trip Factor approach will be the first combination of normalisation strategies explored to test the stated hypothesis. As mentioned in the introduction to the concept of normalisation in section 10.3, Score normalisation is used to attempt to address the simplest kind of bias possible, bias that is consistent across all Descriptors.

The method for determining the Score Bias Normalisation Trip Factor values is presented in section 10.4.1.

10.4.1 Calculation of Score Bias Normalisation by Trip Factor

Let the total of all scores s_{jk} given by practitioner k to subject j in each subgroup be

$$s_{jk} = \sum_{i=1}^{16} s_{ijk}, \qquad (10.1)$$

in which the subscript i refers the diagnostic Descriptor.

Then mean score for each practitioner is then given by

$$\bar{s}_k = \frac{1}{J} \sum_{j=1}^{J} s_{jk}, \qquad (10.2)$$

in which J is the total number of subjects. The mean score for all practitioners becomes

$$\bar{s} = \frac{1}{k}\sum_{k=1}^{K}\bar{s}_k = \frac{1}{JK}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{16}s_{ijk}, \tag{10.3}$$

in which K is the total number of practitioners.

The normalising TPS factor for practitioner k, $F_{k,TPS}$, is then defined as

$$F_{k,TPS} = \frac{\bar{s}}{\bar{s}_k} \tag{10.4}$$

The new "scores" $\left(s_{ijk}\right)_{new_1}$ for each diagnostic characteristic by each practitioner were then obtained from

$$\left(s_{ijk}\right)_{new_1} = s_{ijk}F_{k,TPS} \tag{10.5}$$

so that after normalisation of $\left(s_{ijk}\right)_{new_1}$, the new mean mark for each practitioner would have the same value as with the raw data, namely $\bar{s}$.

10.4.2 Implementation of Score Bias Normalisation by Trip Factor

To calculate the normalisation factors, the scores for each rater on each day were totaled. The total scores of each rater on each day prior to normalisation are presented below in table 10.4.2.1.

Table 10.4.2.1 Sum of Raw scores for each Rater on each day

Raw Scores	Day 1	Day 2	Day 3
Rater 1	373	226	250
Rater 2	275	169	118
Rater 3	368	195	162
Rater 4	283	267	194
Rater 5	491	289	178
Average	358	229.2	180.4

The total scores of each rater from each day and the averages of each day summarised in table 10.4.2.1 were used to calculate the percentage difference each rater's total score was above or below the mean and the results were termed Normalising Factors and are presented in table 10.4.2.2. Instead of using the number determined from equation (10.4) in section 10.4.1, one was subtracted from this value and the result is expressed as a percentage to show the variation from the mean.

Table 10.4.2.2 Normalising Factors of the raters on each day

Difference	Day 1	Day 2	Day 3
Rater 1	4%	-1%	39%
Rater 2	-23%	-26%	-35%
Rater 3	3%	-15%	-10%
Rater 4	-21%	16%	8%
Rater 5	37%	26%	-1%

Superficially had Rater 2 been the same person on the three days then it would appear that he/she always scored low. However, with the same proviso Rater 4 appears to be much less consistent. Different raters were however used each day, so no correlation between the raters scores on each day can be inferred. In any case a question remains: How does one employ

these Normalising Factors to alter a practitioner's score? If scores are simply multiplied by these values, the accuracy of a rating is greatly increased. For example if Rater 2 on Day 1 had entered 3 for a particular Descriptor it would have become 2.31 a change of -0.69 from the original score. Clearly 2.31 is not a number that can be used since the raters were asked to score in integers 0-5.

Since for the same rater on the same day, low scores such as 1 change by much less, viz, - 0.23, and at the other extreme, a score of 5 changes by -1.15, so that inconsistencies occur. Finally the use of two significant figures after the decimal place also may cause the computer code used to evaluate agreement to fail. A process was required that caused scores to be changed in a way, which would more fairly make changes to a rater's score and allow changes to a rater's score by integer values only.

The system adopted was to view scores as 'spheres of influence' or 'addresses' instead of numbers. This caused each score to be treated equally. Thus this approach to removal of bias was named "Trip Factor Normalisation" and involved the application of a threshold to the Normalising Factors.

Since 2.5 is midway between 0 and 5 it was used as the basis for determining the trip factor. In order to change a rating of 2 to 2.51 so that it could be rounded up to 3 a 1.25 multiplier would be required. Similarly in order to change 3 to 2.49 so it could be rounded down to 2, a multiplier of 0.83 would be required. Since the rating can only be changed by an integer the

requirement that a normalising factor less that 0.83 when the value is 3 or greater than 1.25 when the value was 2 would be required. This seemed a rather daunting requirement. Therefore, as a compromise, the strategy adopted was that if a rater had a normalising factor less than or equal to 0.85 the value of each score was reduced by 1 and if the normalising factor was greater than 1.15 the score was increased by 1. If the Normalising Factor was between 0.85 and 1.15 no change was effected to the rater's score.

Factor Normalisation Values and the Normalising Factors again expressed as percentage variations from the mean are presented in Table 10.4.2.3. This of course means that whereas had the original Normalising factors given in Table 10.4.2.2 been used, the averages of the TPS of each rater on each day would have been the same, with the trip factor approach actually adopted, the mean for each practitioner after normalisation is not the mean of all practitioners.

Table 10.4.2.3 Trip Factor Normalisation Values applied to the Raw Data on each day (these are not normalising factors)

		Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Day 1	Normalising Factor	4%	-23%	3%	-21%	37%
	Change in Value	0	-1	0	-1	1
Day 2	Normalising Factor	-1%	-26%	-15%	16%	26%
	Change in Value	0	-1	0[4]	1	1
Day 3	Normalising Factor	39%	-35%	-10%	8%	-1%
	Change in Value	1	-1	0	0	0

[4] The Normalising Factor was rounded two significant figures after the decimal point. The Normalising Factor Rater 3 on day 2's but was just below above 0.85 so the score was not changed.

The Normalising Values from Table 10.4.2.3 were added to the non-zero raw scores and the result capped at five to arrive at the Score Normalised by Trip Factor Scores values.

10.4.3 Score Changes after Score Normalisation by Trip Factor

As mentioned above, the goal of normalisation was to cause the total scores of each of the raters on each day to approach equality. To test whether this outcome occurred in this first attempt at normalisation, the sum of the deviations of the total scores for each rater from the mean of each Descriptor summed over all Descriptors on each day for each practitioner before and after normalisation are presented together in Table 10.4.3.1. Further, since the mean changed its value after normalisation, the fraction of the mean that the sum of the deviations represents expressed as a percentage, is also presented in Table 10.4.3.1.

Table 10.4.3.1 Total from average on each day before and after Score Normalisation by Trip Factor approach implementation

		Day 1		Day 2		Day 3	
Raters		Scores	Difference from mean	Scores	Difference from mean	Scores	Difference from mean
Rater 1	Raw score[1]	15	4%	-3.2	39%	69.6	11%
	Norm score[2]	13.6	4%	16.4	-9%	-7.6	-2%
Rater 2	Raw score	-83	-23%	-60.2	-35%	-62.4	-27%
	Norm score	0.6	0%	27.4	0%	-8.6	0%
Rater 3	Raw score	10	3%	-34.2	-10%	-18.4	-6%
	Norm score	8.6	2%	-14.6	4%	-10.6	3%
Rater 4	Raw score	-75	-21%	37.8	8%	13.6	-3%
	Norm score	5.6	2%	-26.6	7%	21.4	1%
Rater 5	Raw score	133	37%	59.8	-1%	-2.4	25%
	Norm score	-28.4	-8%	-2.6	12%	5.4	-2%

Total score differences from the mean are reduced in nearly all cases after Score Normalisation by Trip Value method, as can be seen in Table 10.4.3.1. The most striking example is on day two, Rater 2, in which the total score difference from the mean changed from -60.2 to 27.4. Since a change in sign of the difference occurred, this example indicates that the normalisation process sometimes over-compensated for the pre-normalisation differences from all raters' mean scores, an undesirable result. It seemed that the Trip Factor approach has to be used carefully since the score adjustments might be too severe.

1 Raw Score is the score value before any Normalisation

2 Norm Score is an abbreviation for the score after Normalisation

Despite this effect, it is clearly seen in Table 10.4.3.2 that after normalisation, the sum of the absolute differences of the ratings of each rater from the mean of all raters on a particular was significantly lower than the same sum on the raw scores.

Table 10.4.3.2 Total absolute differences from the mean in the raw and Score Normalised by Trip value data

	Raw	Normalised
Day 1	316	56.8
Day 2	195.2	87.6
Day 3	166.4	53.6
All Days	677.6	198

The normalised data was sorted into the three Wellness Groups according to the raw TPS of each subject originally presented in Chapter 4, thereby allowing the calculation of inter-rater agreement. The original raw TPS scores were used as the criterion for inclusion of subjects into the wellness groups so that a valid comparison of the subjects in each Wellness Group before and after normalisation can be made. If the TPS of the subjects after normalisation were used as the criterion for inclusion in the Wellness Groups, the subjects included in each group may have been different. Each wellness group has a mixture of subjects from the three different data collection days, so analysis of score changes can only be carried out in each day's data, prior to allocation to the Wellness Groups.

10.4.4 Inter-Rater Agreement with Score Bias Normalisation by Trip Factor

The linearly weighted simple and AC2 agreements with the data obtained after normalisation with the Trip Factor are presented in Table 10.4.3.1.

Table 10.4.4.1 Linearly Weighted simple and AC2 agreement in the Trip Factor Normalised data

Groups	Simple Agreement	AC2
Most Well	0.86 ±0.01	0.79 ±0.02
Intermediate	0.77 ±0.01	0.59 ±0.03
Least Well	0.73 ±0.01	0.44 ±0.03
All Groups	0.78 ±0.01	0.61 ±0.02

These agreements were compared to those obtained with the raw DSOM data and are presented in Table 10.4.4.2. The values were found by subtracting the agreement values obtained with the DSOM data shown in Table 4.4.1 from those in Table 10.4.3.1 and obtaining the root mean square standard error from the standard errors in the two tables.

Table 10.4.4.2 Differences in linearly weighted and AC2 agreements between the raw DSOM and Trip Factor normalised data

Groups	Simple Agreement	AC2
Most Well	0.01 ±0.01	0.02 ±0.03
Intermediate	0.01 ±0.02	0.02 ±0.04
Least Well	0.00 ±0.02	0.02 ±0.04
All Groups	0.00 ±0.01	0.01 ±0.02

Despite the convergence of total scores of the raters on each day after the normalisation, as may be observed in Table 10.4.3.1, the differences between raters before and after normalisation are quite small in both simple and AC2

agreement. Further since the root mean standard errors associated with these differences are larger than the differences, it can be concluded that there is no actual difference in the agreement before and after normalisation.

Trip Factor normalisation is a 'blunt' instrument. If scores changed after normalisation, they were increased or decreased simply by a single integer. If the differences between the raters' scores were large or small, no consideration or accommodation was given within the Trip Factor approach. The Trip Factor approach was obviously ineffective and produced flawed normalisation outcomes.

10.5 Score Bias Normalisation by a "Score Factor"

Another, more elaborate approach would appear to be necessary so that the Score Factor approach was therefore developed. A Score Factor amount was calculated and applied to each non-zero score in order to make the total average scores of all the raters equal. The Score Factor amount was calculated by dividing the sum of differences between the TPS of each rater for each subject on each day from the mean by the sum of the number of non-zero scores allocated by this rater for all subjects on that day.

Again, in the same way as with the Trip Factor approach, there were extra constraints. To enable the AC2 program to process the data, the Score Factor amount was rounded to the nearest 0.5 so as to allow the AC2 program to cope with resulting weighting table. If the matrix of the weights becomes large the program would become very slow and sometimes crashes. For example, if

the AC2 program was used with Score Factor values that were rounded to the nearest 0.1, unexplained reductions in the inter-rater agreement occurred compared with the results obtained with values rounded to the nearest 0.5. The Score Factor values were also capped at ±2 and was not applied to zero scores and finally the normalised score had to be in the range 0-5.

10.5.1 Score Bias Normalisation by Score Value Calculation Equation

Let T_k be the total of all the scores that practitioner k allocated, namely,

$$T_k = \sum_{i=1}^{16} \sum_{j=1}^{J} s_{ijk}. \tag{10.6}$$

The average of the total of scores for all practitioners is therefore

$$\bar{T} = \frac{1}{K} \sum_{k=1}^{K} T_k = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{16} \sum_{j=1}^{J} s_{ijk}. \tag{10.7}$$

A normalising factor F_k for practitioner k can then be written as

$$F_k = \frac{T_k - \bar{T}}{M_k}, \tag{10.8}$$

in which M_k is the number of non-zero scores entered by practitioner k.

However, the range and precision of F_k needs to be controlled so that F_k' the rounded value of F_k to the nearest 0.5 is given by

$$F_k' = sign(F_k) \frac{\lfloor 2 sign(F_k) F_k + 0.5 \rfloor}{2} \tag{10.9}$$

and F_k' is constrained to the range

$$-2 \leq F'_k \leq 2. \qquad (10.10)$$

The modified value of all the scores $(n_{ijk})_{new_2}$ entered by practitioner k become

$$\left(s_{ijk}\right)_{new2} = s_{ijk} + F'_k \qquad (10.11)$$

10.5.2 Score Bias Normalisation by Score Value Calculation

The values used for calculation and the Score Bias Normalisation by Score Values is presented in Tables 10.5.2.1 (a-c).

Column three lists the score differences of each rater, which were divided by the corresponding value in column two, the total Non-zero scores of each rater. This produced the Raw Score Factors located in column four, which was rounded to the nearest 0.5 and capped where necessary to plus or minus two in the far right column. The Rounded Score Factors were applied to the raw scores non-zero to create the Score Normalised by Score Factor Data and the outcomes capped at five and zero as per the guidelines for all normalisation processes set forth in section 10.3.

Table 10.5.2.1a Score Normalisation by Score Value calculation, day one

Rater	1	2	3	4	5
Number of non-zero scores	138	86	135	93	160
Sum of difference between TPS and mean TPS of all Raters	15	-83	10	-75	133
Raw Score Factor	-0.1	1	-0.1	0.8	-0.8
Rounded Score Factors	0	1	0	1	-1

Table 10.5.2.1b Score Normalisation by Score Value calculation, day two

Rater	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Number of non-zero scores	95	68	92	84	82
Sum of difference between TPS and mean TPS of all Raters	-3.2	-60.2	-34.2	37.8	59.8
Raw Score Factor	0	0.9	0.4	-0.5	-0.7
Rounded Score Factors	0	1	0.5	-0.5	-0.5

Table 10.5.2.1c Score Normalisation by Score Value calculation, day three

Rater	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Number of non-zero scores	85	47	64	72	69
Sum of difference between TPS and mean TPS of all Raters	69.6	-62.4	-18.4	13.6	-2.4
Raw Score Factor	-0.8	1.3	0.3	-0.2	0
Rounded Score Factors	-1	1.5	0.5	0	0

Whilst at first glance this normalising technique should have resulted in more appropriate adjustments to raters' score than the Trip Factor approach, in fact, as may be seen in Table 10.5.2.2, this was not the case. Whilst eight adjustments to the raters' scores were made with the Trip Factor approach, ten were made with the Score Value approach. Two of the raters whose scores required no adjustment when the Trip factor was used had adjustments of only 0.5 each in the case of Amount. Indeed, the level of subtlety of score change was also greater when Score Factor Normalisation was used; there are two occasions where the score each of the raters scores was altered by only -0.5 when the Score Factor approach was employed whereas with the Trip Factor approach, it had been -1 since when the Trip Factor approach had been used only changes of ± 1 were permitted. There

are one occasion where the amount Normalising was greater than one, another variation from the trip factor method.

Surprisingly, there is little difference between the two methods in the values used to “correct” the scores, as is seen in Table 10.5.2.2. The first day’s normalising factors were exactly the same with both methods. The two normalising factors were exactly the same on ten occasions in total and in the other five instances each value was within 0.5.

Table 10.5.2.2 Trip and Score Factors for Score Normalisation for each day and each rater

Days	Day		Day 2		Day 3	
Norm Factor	Trip Factor	Score factor	Trip Factor	Score Factor	Trip Factor	Score Factor
Rater 1	0	0	0	0	-1	-1
Rater 2	1	1	1	1	1	1.5
Rater 3	0	0	0	0.5	0	0.5
Rater 4	1	1	-1	-0.5	0	0
Rater 5	-1	-1	-1	-0.5	0	0

10.5.3 Score changes after Score Normalisation by Score Factor approach

As with the Trip value normalisation approach outcome described in table 10.4.2.4, the day-by-day percentage of difference change of each rater’s total scores before and after normalisation was compared. This comparison is presented in table 10.5.3.1

Table 10.5.3.1 Score Differences from average on each day before and after Score Normalisation by Score Factor approach implementation

		Day 1	Day 2	Day 3
Rater 1	raw	15	-3.2	69.6
	norm	13.6	16.4	-7.6
Rater 2	raw	-83	-60.2	-62.4
	norm	0.6	27.4	-8.6
Rater 3	raw	10	-34.2	-18.4
	norm	8.6	-14.6	-10.6
Rater 4	raw	-75	37.8	13.6
	norm	5.6	-26.6	21.4
Rater 5	raw	133	59.8	-2.4
	norm	-28.4	-2.6	5.4

As would be expected from the factor comparison made in Table 10.5.2.2, the differences between the sum of the TPS given by each rater and the mean value of these sum for all raters are reduced by similar amounts as when Trip Factor Normalisation is used. A way to demonstrate the “global” effect of the change in scores after Amount Normalisation is to show that there is a reduction in the sum of the deviations of the total of all TPS given by all rater on a given day from the mean of all sums of the TPS of all the raters on that day. As may be seen in Table 10.5.3.2, the differences from the mean decreased markedly after Amount Normalisation. The magnitude of the decrease from its value before Normalisation ranges from a minimum 75% on the third day to maximum of 82% on the first day with a mean of 80% over the three days; a very significant reduction. This suggested that there might be a significant improvement in the inter-rater agreement. The same three wellness groups were used to evaluate the inter-rater agreement.

Table 10.5.3.2 Total absolute differences from the mean in the raw and Score Normalised by Score value data

	Raw	Normalised	Percentage change
Day 1	316	56.8	82
Day 2	195.2	39.6	80
Day 3	166.4	41	75
All Days	677.6	137.4	80

10.5.4 Inter-Rater Agreement after Score Factor Normalisation

The inter-rater simple and AC2 linearly weighted agreements after the DSOM data had been normalised with the Score Bias by Score Value approach are presented in Table 10.5.4.1.

Table 10.5.4.1 Linearly weighted simple and AC2 agreement after Score Normalisation by Score Factor in the three Wellness Groups and All Groups

Groups	Simple Agreement	AC2
Most Well	0.83 ±0.01	0.75 ±0.02
Intermediate	0.76 ±0.01	0.58 ±0.03
Least Well	0.73 ±0.01	0.44 ±0.03
All Groups	0.78 ±0.01	0.61 ±0.02

As may be seen in Table 10.5.4.2, when these agreement data are compared with agreements obtained with the raw data found in table 4.4.1 and the differences between the agreements and the root mean square standard error from the standard errors are calculated from the data of the two tables, a similar outcome is observed as that which occurred when agreements obtained with the Score Normalised by Trip Factor data were compared with the agreements obtained with the raw DSOM data.

Table 10.5.3.2 Differences between Score Normalised by Score Factor and Raw data agreement

Groups	Simple Agreement	Standard Error	AC2	Standard Error
Most Well	0.01	0.01	-0.02	0.03
Intermediate	0.00	0.02	0.02	0.04
Least Well	0.00	0.02	0.02	0.05
All Groups	0.00	0.01	0.01	0.02

The differences in both the linearly weighted simple and chance-removed AC2 agreements in all Wellness Groups and All Groups combined before and after normalisation are negligibly small with a very small error.

After successfully causing the average scores of each rater to be much closer to the mean by the application of the two approaches used to attempt to remove bias, no significant improvements in agreement were obtained. This suggests no rater consistently scored high or low that is on the average no rater was consistently biased.

10.6 Descriptor Bias Normalisation by "Trip Factor"

The results obtained with Score Bias normalisation by the two approaches utilised do not completely rule out the possibility of raters being consistently biased; the raters' bias may be consistent within a particular Descriptor, but perhaps be different from other Descriptors scored by the same rater.

The possibility of consistent bias unique to Descriptors could explain the lack of improvement in diagnostic agreements after implementations of the Trip

Factor and Score Factor approaches to Score Normalisation. To explore this possibility, each Descriptor had to be examined separately. In the first instance, the simpler Trip Factor approach was utilised to normalise the data of each individual Descriptor.

10.6.1 Descriptor Bias Normalisation by Trip Factor Calculation Equation

Here the mean of each of the Diagnostic Descriptors for each practitioner were adjusted to have the same value. The mean, $\bar{s}_{ki}$, for each practitioner, k, for each diagnostic Descriptors i, is given by

$$\bar{s}_{ki} = \frac{1}{J}\sum_{j=1}^{J} s_{ijk}. \qquad (10.12)$$

The mean mark for all practitioners for Diagnostic Descriptor, i, $\bar{s}_i$ is given by

$$\bar{s}_i = \frac{1}{K}\sum_{k=1}^{K} \bar{s}_{ki} = \frac{1}{JK}\sum_{k=1}^{K}\sum_{j=1}^{J} s_{ijk}. \qquad (10.13)$$

The normalising factor for practitioner, k, and Diagnostic Descriptor, i, now becomes

$$F_{ik} = \frac{\bar{s}_k}{\bar{s}_{ki}} \qquad (10.14)$$

and the new score becomes,

$$\left(s_{ijk}\right)_{new_3} = s_{ijk} F_{ik} \qquad (10.15)$$

so that the mean for each of the Diagnostic Descriptors for each of the practitioners becomes $\overline{s}_i$.

Clearly since equation (10.4) is different from equation (10.13) different normalised values are obtained which need to be interpreted differently. However, once again the values of $\left(s_{ijk}\right)_{new_3}$ will be unlikely to be integers, so that similar approaches to those used in section 10.4 will need to be adopted.

10.6.2 Descriptor Bias Normalisation by Trip Factor Calculation

The same process as was employed in section 10.4 to obtain integer values after normalisation is adopted here. The trip factors for the total scores for all Descriptors was developed and applied in exactly the same way as in Section 10.4, except in the current process each Descriptor was normalised separately. As already set forth in Chapter 10.4, the foundation of the process Trip Factor Normalisation was the percentage differences of each rater from the mean of all raters. The total scores in each Descriptor and rater and the average scores of each Descriptor have already been presented in tables 4.6.2.1(a-c) and discussed in section 4.6.

The differences in raters' total scores will be represented in tables presented that calculate the percentage differences each rater's score varied from the mean of all raters. The critical information which the Trip factor normalisation approach used to determine which scores would be adjusted and the adjustments made to the non-zero raw scores are reported next to these percentage differences. As in Score Normalisation by the Trip Factor

approach, when the absolute values of these differences were greater or less than 15% from the mean of all raters, a Trip Factor and the raw scores were adjusted by ±1 to become the normalised scores. Rater-Descriptor totals if zero are also clearly denoted with 'No score'. These data are presented in Appendix 9(a-c). The number of times the normalisation process caused each Descriptor to be altered up or down for each rater on each day is an indication of each rater's scoring bias. In some cases, a rater's scores are almost consistently normalised in one direction, which could indicate a consistent overall bias. Raters two, four and five on day one, raters one and five on day two and rater one on day three tended to almost exclusively have adjustments in one direction. On the other hand, some, for example rater 3 on day one, rater two on day two and rater five on day three had roughly equal proportions of up and down adjustments. This meant that some raters' scores were consistently high or low, but other raters varied in scoring tendencies in individual Descriptors. It is impossible to detect the fact that all raters may have scored consistently high or low since implicitly it has been assumed that the average of the raters scores is the "true" score. Unless there is a method of determining the "correct" diagnosis such as a conclusive physical measurement that can be used, there is no way of determining whether the consensus diagnosis, that is the mean of all the raters, which by default becomes the diagnosis is indeed the correct diagnosis.

The variations observed in scoring, with some cases raters scoring higher than average in some Descriptors but lower than average in others means that Descriptor Normalisation could provide different agreement outcomes

than occurred after Score Normalisation and therefore had the potential to improve inter-rater agreement.

The percentage differences from average in the Descriptors, after each Descriptor was normalised individually, as presented in Appendix 9(a-c) are much higher than those of the table 10.4.2.2 in which Score Normalisation is presented. This is particularly apparent on day three when only eight subjects were interviewed. Fewer subjects on a particular day seemed to be associated with greater fluctuations in differences between raters. This is logical, as there are smaller numbers to provide a reliable average. The application of the normalising factors led to score changes to the appropriate Descriptor-rater's non-zero scores and the score changes were capped to the maximum of five as well.

10.6.3 Scores and Score changes after Descriptor Normalisation by the Trip Factor approach

The scores of each rater-Descriptor combination on each day after Descriptor Normalisation by Trip Factor approach are presented in Appendix 10(a-c) Scores in each rater-Descriptor combination after Descriptor Normalisation by Trip Factor application. It is difficult to determine the overall effects of Descriptor Normalisation by Trip Factor from the data presented in Appendix 10(a-c), due to the large quantity of data. These results need to be compared to data prior to normalisation presented in Table 4.6(a-c) to have meaning. The differences between the scores in each rater-Descriptor combination before and after Trip Factor normalisation are easier to understand when

viewed as differences between the Descriptor-rater total scores from the mean before and after applying the Trip Factor Normalisation process. These data are presented in Appendix 11(a-c) Differences from the mean of Descriptor-rater totals in Raw and Descriptor Normalised by Trip factor method data, days one to three.

An important observation that only became apparent during the implementation of both attempts of Descriptor normalisation, was that some of the raters provided scores of zero in certain Descriptors for all subjects they interviewed. It seems that they did not believe that that Descriptor was a legitimate disease state that they believed any person could have. This meant that under the rules previously defined that were used to conduct normalisation these Descriptor-rater combinations did not receive any score changes. The Descriptor Not Scored was identified with the abbreviation of NS in the relevant tables.

It is interesting which Descriptors were not scored or was not very often scored, that the sum of all scores was low. For example Phlegm had no scores from one rater on day 1 and from four out of the five raters on day 3 with two raters giving very low scores on day 2. Dry had no scores from raters on days one and two. Cold and Yang Xu had no scores from one rater and on days 2 and 3. Blood stagnation had no scores from one rater on day two. There were in total 11 un-scored rater-Descriptors combinations from a total of 240 Descriptor-rater combinations on the three days and the Descriptors which were un-scored by raters naturally had very high agreement.

Similarly, high levels of agreement were experienced with Descriptors that were only scored above zero spasmodically. The observation of such low scoring of the Phlegm and Dry Descriptors confirms their deletion in the CMDD format was an appropriate decision.

Generally however, the Descriptor Normalisation by Trip Factor approach seemed to work moderately well. However, due to the process of normalising data in much smaller volumes in the single Descriptor-rater day groups than in the All Descriptor-rater groups used for the Score Normalisations, there were occasionally untoward and unwanted outcomes after the implementation of the Descriptor Normalisation by Trip Factor approach.

Sometimes what could be called over-runs occurred in Descriptor-rater score totals, after the Trip Factor implementation, which describe a change in the sign of the change in the difference of the scores of a rater from the average in a Descriptor-rater total, while in other cases no changes after normalisation were observed. While in many cases these over-runs were minor and only moved to a minor degree to the opposite polarity and were therefore deemed insignificant, there were a few cases where a Descriptor-rater total deviated excessively in one direction prior to normalisation which were reversed strongly to the opposite polarity. This outcome is clearly against the spirit of the normalisation process; is not a desirable effect that may have corrupted the data for inter-rater agreement calculations.

This over-run event beyond where the average score was greater than five in each direction from the mean occurred in a total of six occasions and while this is a minor number of occurrences compared to the total of Descriptor-rater day combinations of eighty each day, or two hundred and forty in total, it is nonetheless an undesirable outcome.

Many of the Descriptor-rater score differences were below the Trip Factor of ±15% and were therefore left unaffected. There were 23 instances of Descriptor-rater combinations that had no score change on day one, 20 such instances on day two and 28 on day three. When the normalisation over-runs and the effects of no score change outcomes are combined, a picture of an unsatisfactory implementation comes to light, with a total of 88 of 240 Descriptor-rater combinations either not normalised or incorrectly so.

The goal of the normalisation processes discussed above, was to cause the overall difference from the mean to be reduced. This is confirmed by an inspection of Table 10.6.3.1, in which the reduction in the sum of the average absolute differences of all raters from the average of all raters between the raw and normalised scores on each day is presented.

Table 10.6.3.1 Total absolute differences from the mean in raw and normalised data on each day

	Raw	Normalised
Day 1	560.4	263.6
Day 2	396.8	186.8
Day 3	320.0	176.0

10.6.4 Results of Descriptor Bias Normalisation by Trip Factor

Agreement was calculated in each Descriptor after Trip Factor normalisation. As well as overall agreement in all subjects in each Descriptor, agreement was also calculated in three intra-Descriptor Wellness Groups within each Descriptor.

Agreement after normalisation was compared to the results reported in Chapter 4.7, where the same intra-Descriptor-subject groupings were investigated within the raw DSOM data, with the subject groupings again decided according to each subject's raw Total Practitioner Score to maintain consistency of subject groupings and agreement calculated with Linearly weighted simple agreement and the AC2 statistic

Table 10.6.3.1a Linearly Weighted Simple Agreement after Descriptor Normalisation by Trip Factor approach

Groups	Least Well		Intermediate		Most Well	
Statistic	Simple Agreement	Std Error	Simple Agreement	Std Error	Simple Agreement	Std Error
Blood Stag	N/A		0.72	0.03	0.89	0.02
Blood Xu	N/A		0.60	0.03	0.88	0.02
Cold	N/A		0.61	0.03	0.91	0.02
Damp	N/A		0.63	0.03	0.90	0.02
Dry	N/A		0.63	0.02	0.90	0.02
Heart	0.81	0.06	0.58	0.04	0.90	0.03
Heat	0.92	0.05	0.60	0.02	0.60	0.02
Kidney	0.75	0.04	0.55	0.02	0.80	0.03
Liver	0.74	0.04	0.60	0.02	0.86	0.04

Table 10.6.3.1a Linearly Weighted Simple Agreement after Descriptor Normalisation by Trip Factor approach (continued)

Lung	0.82	0.02	0.62	0.03	0.89	0.02
Phlegm	N/A		0.58	0.05	0.90	0.02
Qi Stag	0.71	0.15	0.63	0.02	0.80	0.03
Qi Xu	0.77	0.04	0.64	0.03	0.81	0.03
Spleen	0.70	0.04	0.65	0.02	0.80	0.03
Yang Xu	N/A		0.60	0.02	0.92	0.02
Yin Xu	0.72	0.04	0.62	0.04	0.87	0.03
Average	0.77	0.05	0.62	0.03	0.85	0.02

Table 10.6.3.1b Linearly Weighted AC2 Agreement after Descriptor Normalisation by Trip Factor approach

Groups	Least Well		Intermediate		Most Well	
Statistic	AC2	Std Err	AC2	Std Err	AC2	Std Err
Blood Stag	N/A		0.41	0.06	0.87	0.03
Blood Xu	N/A		0.12	0.06	0.85	0.03
Cold	N/A		0.19	0.09	0.89	0.03
Damp	N/A		0.14	0.09	0.88	0.03
Dry	N/A		0.27	0.04	0.94	0.02
Heart	0.65	0.15	0.03	0.08	0.88	0.04
Heat	0.87	0.09	0.08	0.06	0.82	0.05
Kidney	0.48	0.08	0.06	0.04	0.73	0.06
Liver	0.48	0.10	0.06	0.03	0.80	0.07
Lung	0.67	0.04	0.16	0.08	0.87	0.03
Phlegm	N/A		0.19	0.10	0.88	0.03
Qi Stag	0.45	0.15	0.10	0.15	0.71	0.06
Qi Xu	0.52	0.10	0.14	0.07	0.72	0.06
Spleen	0.37	0.10	0.16	0.05	0.66	0.06
Yang Xu	N/A		0.22	0.06	0.90	0.03
Yin Xu	0.37	0.11	0.16	0.10	0.84	0.04
Average	0.54	0.10	0.16	0.07	0.83	0.04

The data presented in tables 10.6.3.1a and b are difficult to interpret without direct comparison with the agreement results obtained with the raw DSOM data agreement results. Therefore, the differences between the raw and normalised agreement values and the root mean squares of the standard errors are presented in tables 10.6.3.2a and b. This format makes effective analysis and discussion of the data possible.

Table 10.6.3.2a Changes in Linearly Weighted Simple Agreement and standard errors after Descriptor Normalisation by Trip Factor approach

Difference	Least Well		Intermediate		Most Well	
Statistic	Simple Agree	Std Err	Simple Agree	Std Err	Simple Agree	Std Err
Blood Stag	N/A		0.08	0.04	0.05	0.03
Blood Xu	N/A		0.02	0.05	0.00	0.03
Cold	N/A		0.02	0.04	0.02	0.03
Damp	N/A		-0.01	0.04	0.01	0.03
Dry	N/A		0.02	0.03	-0.02	0.03
Heart	0.08	0.07	0.01	0.05	0.00	0.04
Heat	0.15	0.09	-0.04	0.03	-0.27	0.03
Kidney	0.01	0.05	-0.01	0.03	-0.01	0.04
Liver	0.01	0.05	-0.01	0.02	-0.01	0.05
Lung	0.04	0.06	-0.02	0.03	-0.03	0.03
Phlegm	N/A		0.01	0.05	-0.01	0.03
Qi Stag	-0.04	0.16	-0.01	0.02	-0.02	0.04
Qi Xu	0.02	0.05	-0.02	0.04	-0.02	0.04
Spleen	-0.13	0.05	-0.01	0.03	0.01	0.04
Yang Xu	N/A		0.04	0.03	0.02	0.03
Yin Xu	-0.07	0.08	0.02	0.05	0.02	0.04
Average	0.01	0.07	0.01	0.04	-0.02	0.03

Table 10.6.3.2b Changes in Linearly Weighted AC2 Agreement after Descriptor Normalisation by Trip Factor approach

Difference	Least Well		Intermediate		Most Well	
Statistic	AC2	Std Err	AC2	Std Err	AC2	Std Err
Blood Stag	N/A		0.24	0.08	0.06	0.05
Blood Xu	N/A		0.08	0.09	0.01	0.05
Cold	N/A		0.05	0.12	0.03	0.04
Damp	N/A		-0.03	0.11	0.02	0.04
Dry	N/A		0.07	0.05	0.03	0.03
Heart	0.23	0.20	0.06	0.11	0.00	0.05
Heat	0.35	0.19	-0.11	0.07	0.00	0.07
Kidney	0.02	0.13	0.01	0.07	-0.01	0.07
Liver	0.06	0.14	-0.05	0.07	-0.02	0.09
Lung	0.09	0.15	-0.08	0.09	-0.03	0.04
Phlegm	N/A		-0.04	0.11	-0.02	0.04
Qi Stag	-0.04	0.21	-0.04	0.16	-0.02	0.08
Qi Xu	0.02	0.13	-0.05	0.10	-0.03	0.08
Spleen	0.11	0.16	-0.03	0.08	0.02	0.10
Yang Xu	N/A		0.15	0.09	0.03	0.04
Yin Xu	-0.20	0.23	0.07	0.12	0.04	0.05
Average	0.07	0.17	0.02	0.09	0.01	0.06

In almost all cases, the combined standard errors are much greater than the differences observed. It seemed that as there were score total over-runs in these Descriptors after normalisation, the normalisation process did not achieve the goal of causing each rater's score to approach the mean of all scores.

The last attempt to obtain definite inter-rater agreement improvements after the application of normalisation was the Descriptor Normalisation by Score Value approach.

10.7 Descriptor Bias Normalisation by Score Factor

Score Bias Normalisation by Score Factor is similar to the method proposed in section 10.5. Score factors were calculated for each Descriptor-rater combination of each day and these factors were then applied to the non-zero scores.

The Score Factor normalisation approach largely prevents normalisation over-runs and also decreases the number of score non-changes that were both observed in the Descriptor Normalisation by Trip Factor method.

10.7.1 Equation for Descriptor Bias Normalisation by Score Factor Calculation

The mean score given to subjects for Descriptor i by practitioner k subjects is given in Equation (10.12) and the mean score all practitioners gave for Descriptor i is given in equation (10.13). The Score Factor for Descriptor i and practitioner k can be calculated from

$$F_{ik} = \frac{\bar{S}_{ik} - \bar{S}_i}{M_{ik}}, \qquad (10.16)$$

$$F_{ik} = \frac{T_{ik} - \bar{T}_i}{M_i} \qquad (10.17)$$

in which M_{ik} is the number of non-zero scores entered for Descriptor i by practitioner k.

As before, the range and precision of F_{ik} needs to be controlled so that F'_{ik} the rounded value of F_{ik} to the nearest 0.5 is again given by Equation (10.14) rewritten as

$$F'_{ik} = sign(F_{ik}) \frac{\lfloor 2sign(F_{ik})F_{ik} + 0.5 \rfloor}{2} \qquad (10.18)$$

and F'_{ik} is once again constrained to the range

$$-2 \le F'_{ik} \le 2. \qquad (10.19)$$

The modified value of all the scores $(n_{ijk})_{new_4}$ entered by practitioner k become

$$(n_{ijk})_{new_4} = n_{ijk} + F'_{ik} \qquad (10.20)$$

10.7.2 Values calculated for Descriptor Bias Normalisation by Score Factors

The same methods applied in the Score Normalisation by Score Factor were used in Descriptor Normalisation. The Score Factor processes for normalisation were applied to each individual Descriptor separately. Tables for the calculation of score differences from the mean and the number of non-zero scores in each rater-Descriptor combination that were used to calculate the Score Factors for Descriptor Normalisation are presented in the appendices 12(a-c) and 13(a-c) due to the large amounts of data involved.

The differences between each rater-Descriptor combination and the mean of all five rater-Descriptors presented in Appendices 12(a-c) was divided by the number of non-zero scores made in the same rater-Descriptor combination presented in Appendices 13(a-c), to arrive at the raw Score Factor and rounded and capped normalisation values which are presented in Appendices 14(a-c)

There were 29 raw Score Normalisations that exceeded the ±2 capping constraint, nine on day one, eight on day two and twelve on day three. In Blood Xu on day one and in Qi Stagnation on day three, raw Score Normalisation factors even exceeded ten, indicating that there were quite large differences between raters' scores in those Descriptors.

A comparison between the normalising values applied to the raw scores from the two methods used to implement Descriptor Normalisation is made in Appendices 15(a-c).

The score differences between the two different types of normalisation were at most ±1 and more often ±0.5. The tables presented in Appendices 15(a-c) are difficult to comprehend; as there is much data, so Table 10.7.2.1 was created, which summarises the frequency of each level of difference.

Table 10.7.2.1 Summary of differences between trip factor and score factor normalisation values applied to implement Descriptor Normalisation

Differences	Count	Percentage
±0.5	97	41%
±1	39	16%
No difference	104	43%
Totals	240	100%

Only sixteen percent of the normalisation values differed by ±1, the vast majority, over eighty percent of the normalisation values were either exactly the same or within a ±0.5 difference.

There were also more rater-Descriptor combinations adjusted by the Score Value normalisation than in the Trip Factor approach. Table 10.7.2.2 provides the observation that there was around a third reduction between the former and latter methods.

Table 10.7.2.2 Non-normalisations that occurred each day and overall in the Trip and Score Factor approaches

	Trip Factor	Score Factor
1	21	14
2	18	11
3	22	14
All	61	39

In the next section, the changes in scores after normalisation are presented.

10.7.3 Score Changes after Descriptor Normalisation by "Score Factor"

The Capped Score Factors were next applied to the raw scores and the totals constrained where necessary to retain the initial range of 0-5. The total rater-

Descriptor scores after the score factors were applied on each day are presented in the appendix table 15(a-c). The total score changes in each rater-Descriptor combination after the application of the Score Factors is presented in appendix table 16(a-c).

These data in appendices 15 and 16 are difficult to interpret, so the average absolute differences of each rater-Descriptor combination from average score were added and the totals presented in Table 10.7.3.1. An overall convergence to the mean after Descriptor by Score Value normalisation application was observed in this table.

Table 10.7.3.1 Absolute total score differences from the mean of the raw and Descriptor normalised by Score Factor data

	Raw	Trip Factor Normalised	Score Factor Normalised
Day 1	560.4	263.6	181.2
Day 2	396.8	186.8	105.6
Day 3	320	176.0	98.8

The table confirms that the normalisation process did cause the total scores of each rater to significantly converge and approach the mean.

10.7.4 Agreements after Descriptor Normalisation by Score Factor

With the same structure as the analysis of agreement in Descriptor Normalisation by Trip Factor investigation reported in 10.6, as well as overall agreement in all subjects in each Descriptor, agreement was also calculated in three intra-Descriptor Wellness Groups within each Descriptor.

The linearly weighted simple and AC2 inter-rater agreement after normalisation was compared with the results presented in Section 4.7, in which the same intra-Descriptor-subject groupings were employed in the study with the raw DSOM data. The results are presented in table 10.7.4.1 and 10.7.4.2.

Table 10.7.4.1 Linearly Weighted Simple Agreement after Score Factor Normalisation Implementation

Groups	Least Well		Intermediate		Most Well	
	Simple Agreement	Std Error	Simple Agreement	Std Error	Simple Agreement	Std Error
Blood Stag	N/A		0.71	0.02	0.91	0.02
Blood Xu	N/A		0.59	0.03	0.90	0.02
Cold	N/A		0.62	0.03	0.92	0.02
Damp	N/A		0.64	0.03	0.91	0.02
Dry	N/A		0.66	0.02	0.91	0.02
Heart	0.77	0.04	0.61	0.04	0.91	0.02
Heat	0.72	0.10	0.64	0.02	0.64	0.02
Kidney	0.72	0.03	0.59	0.02	0.82	0.03
Liver	0.73	0.04	0.61	0.02	0.88	0.03
Lung	0.77	0.03	0.63	0.04	0.92	0.02
Phlegm	N/A		0.59	0.04	0.93	0.02
Qi Stag	0.74	0.03	0.63	0.02	0.84	0.03
Qi Xu	0.77	0.03	0.67	0.03	0.82	0.03
Spleen	0.68	0.03	0.66	0.02	0.82	0.02
Yang Xu	N/A		0.66	0.03	0.92	0.02
Yin Xu	0.71	0.04	0.57	0.02	0.88	0.02
Average	0.73	0.04	0.63	0.03	0.87	0.02

Table 10.7.4.2 Linearly Weighted AC2 Agreement after Score Factor Normalisation Implementation

Groups	Least Well		Intermediate		Most Well	
	AC2	Std Error	AC2	Std Error	AC2	Std Error
Blood Stag	N/A		0.31	0.05	0.89	0.03
Blood Xu	N/A		0.06	0.05	0.87	0.03
Cold	N/A		0.12	0.09	0.91	0.03
Damp	N/A		0.19	0.09	0.88	0.03
Dry	N/A		0.27	0.05	0.94	0.02
Heart	0.48	0.08	0.04	0.07	0.90	0.03
Heat	0.35	0.25	0.13	0.06	0.84	0.04
Kidney	0.29	0.03	0.10	0.06	0.75	0.05
Liver	0.33	0.11	0.06	0.05	0.84	0.06
Lung	0.46	0.07	0.16	0.11	0.91	0.03
Phlegm	N/A		0.17	0.10	0.91	0.02
Qi Stag	0.34	0.11	0.09	0.06	0.75	0.05
Qi Xu	0.43	0.11	0.19	0.06	0.72	0.06
Spleen	0.18	0.10	0.14	0.05	0.68	0.06
Yang Xu	N/A		0.25	0.07	0.90	0.03
Yin Xu	0.30	0.13	0.06	0.08	0.84	0.03
Average	0.35	0.11	0.15	0.07	0.84	0.04

The data presented in tables 10.7.1.1 and 10.7.1.2 are difficult to interpret without comparison to the non-normalised results. Therefore, the changes in the statistical values from the raw values are next presented in tables 10.7.4.3 and 10.7.4.4; this format makes effective analysis and discussion of the data possible.

Table 10.7.4.3 Changes in Linearly Weighted Simple Agreement after Score Factor Normalisation Implementation

Difference	Least Well		Intermediate		Most Well	
	Simple Agreement	Std Error	Simple Agreement	Std Error	Simple Agreement	Std Error
Blood Stag	N/A		0.06	0.03	0.06	0.03
Blood Xu	N/A		0.01	0.04	0.02	0.03
Cold	N/A		0.03	0.04	0.03	0.03
Damp	N/A		0.00	0.04	0.01	0.03
Dry	N/A		0.05	0.03	-0.01	0.03
Heart	0.04	0.05	0.04	0.05	0.01	0.03
Heat	-0.05	0.12	-0.01	0.03	-0.23	0.04
Kidney	-0.02	0.05	0.03	0.03	0.01	0.04
Liver	-0.01	0.05	0.00	0.02	0.01	0.05
Lung	-0.02	0.06	-0.01	0.05	0.00	0.03
Phlegm	N/A		0.02	0.05	0.01	0.02
Qi Stag	-0.02	0.06	0.00	0.02	0.02	0.04
Qi Xu	0.02	0.04	0.02	0.03	-0.01	0.04
Spleen	-0.16	0.05	0.00	0.03	0.03	0.04
Yang Xu	N/A		0.09	0.04	0.02	0.03
Yin Xu	-0.08	0.08	-0.02	0.04	0.03	0.03
Average	-0.03	0.06	0.02	0.04	0.00	0.03

Table 10.7.4.4 Changes in Linearly Weighted AC2 Agreement after Score Factor Normalisation Implementation

Groups	Least Well		Intermediate		Most Well	
	AC2	Std Err	AC2	Std Err	AC2	Std Err
Blood Stag	N/A		0.14	0.07	0.08	0.05
Blood Xu	N/A		0.03	0.08	0.02	0.05
Cold	N/A		-0.02	0.13	0.04	0.04
Damp	N/A		0.02	0.11	0.02	0.04
Dry	N/A		0.07	0.06	0.03	0.03
Heart	0.05	0.15	0.06	0.10	0.01	0.05
Heat	-0.18	0.30	-0.06	0.07	0.02	0.07
Kidney	-0.17	0.11	0.05	0.08	0.02	0.07
Liver	-0.08	0.15	-0.05	0.07	0.02	0.09
Lung	-0.12	0.16	-0.08	0.12	0.00	0.04
Phlegm	N/A		-0.07	0.12	0.02	0.03
Qi Stag	-0.15	0.24	-0.05	0.07	0.03	0.08
Qi Xu	-0.07	0.18	0.00	0.09	-0.03	0.09
Spleen	-0.08	0.22	-0.05	0.08	0.04	0.10
Yang Xu	N/A		0.18	0.09	0.03	0.04
Yin Xu	-0.27	0.33	-0.03	0.10	0.04	0.05
Average	-0.12	0.21	0.01	0.09	0.02	0.06

The Descriptor Normalisation by Score Factor approach, in a similar fashion to Descriptor Normalisation by Trip Factor again created mild improvement in inter-rater agreement outcomes for all the Descriptor wellness groups. Again, similar to the Trip factor normalisation outcome, any improvements in agreements were almost always within the combined standard errors.

It seems that there was a large component of randomness within the differences between the raters in each case and bias was not consistent, reflected in the lack of inter-rater agreement improvement after the two attempts at normalizations.

The fact that the attempted removal of Score or Descriptor bias by either the Trip Factor or Score factor approaches did not definitely improve agreement means that the hypothesis of practitioner bias being involved as a factor that decreases inter-rater reliability is rejected.

10.8 Normalisation attempts conclusions

It seems that in spite of the many attempts that have been made in this chapter to apply factors to the scores of the raters so as to bring their total scores closer to the mean, there was no definite improvement in agreement obtained. This was not anticipated, as it was reported by other researchers[66] that raters' scores would, according to bias theory[108], contain at least some consistent bias.

It seems then that, as the attempts that were made to remove bias were unsuccessful, that the differences between the raters' choices, were predominantly random and that training designed specifically to remove bias would not lead to significant gains in inter-rater agreement. This finding may be of use in determining the nature of the practitioner training approaches that may be most successful in improving diagnostic accord.

The ability of raters to agree upon the presence or absence of pathology within Descriptors however should not be overlooked and forms a sound basis for effective investigation of treatment interventions or objective phenomena in suitably CMDD diagnosed subjects.

Chapter 11 Conclusion

Methods of determining diagnostic agreement between practitioners after removal of agreement, which would have occurred by chance, have been studied in this thesis. Results of inter-rater diagnostic agreement studies in the literature generally cannot be confidently accepted as correct, since in many cases the statistics used have not been given and in others flawed statistics have been used. The most commonly used approach, when there are more than two raters is Fleiss' Kappa. However, without the diagnostic data being uniformly distributed between all the choices available to raters, termed fixed marginal data, Fleiss' Kappa is an inappropriate and misleading statistical measure of agreement between raters with agreement by chance removed.

It is unlikely that data would be uniformly distributed between all the choices without subjects having been objectively pre-diagnosed and allocated to groups in equal numbers. This appears to be an unrealistic, onerous and certainly undesirable requirement. Despite the fact that this condition has been known for many years, Fleiss' Kappa continues to be used by many researchers. It is emphasised in this thesis that, unless a researcher is certain that the data is fixed marginal, Fleiss' Kappa cannot be used in future studies to evaluate diagnostic agreement between raters.

On the other hand, it is shown that the AC1 statistic can be safely used to determine inter-rater chance-removed agreement when fixed marginal data is

neither expected nor guaranteed. Similarly, in the case of ordinal data, the AC2 statistic should be used since it makes allowances for the proximity of scoring choices between raters in the ordinal data whereas a minor difference in score is treated as a disagreement in determining the AC1 statistic. Further, in the AC2 statistic, different weightings can be used to control the effect of evaluating the effect of the proximity of the scores.

Experiments involving 35 subjects drawn from an open population were performed to evaluate inter-rater agreement between two or three experienced Chinese Medicine practitioners using the traditional Chinese medical format. The simple linearly weighted agreement was low at 19%. The agreement by chance could not be determined in this case, but the simple agreement is well bellow acceptable level. The fact that the traditional Chinese medicine format offers over one hundred diagnostic options that often employ almost, but not exactly the same words to describe variants of the same disease state, which militates against having high agreement.

Further, diagnostic complexity is added by the practice of using an unrestricted number, but generally a patient's state of health is described by two or three diagnoses that are also labeled as to the level of their severity. Unless there is an exact agreement between the words used by raters, all statistical approaches usually used to determine agreement would indicate that there is no agreement. This, therefore, is a major difficulty in evaluating agreement with traditional CM diagnostic format.

It would be very difficult to attempt to investigate the effectiveness of CM treatments, unless one can be at least reasonably sure that the correct diagnosis has been made by the practitioners involved, which seems unlikely with the level of agreement found in the present work. Thus the results of larger scale studies currently planned or being undertaken may not yield reliable results. The potential of 'big data' analysis that may occur in the near future in CM depends on the quality of the practitioners' diagnostic data. It follows that there is an urgent need for acceptable levels of diagnostic consensus between practitioners. A new CM diagnostic format that allows for the differences in diagnosis to be taken into account without destroying the subtlety of the diagnosis and which also allow the application of chance-removed statistics is therefore required.

A diagnostic format termed "Diagnostic System of Oriental Medicine" (DSOM) was identified as having the potential to improve agreement exceeding that found with the contemporary CM diagnostic format. An experiment involving 42 subjects drawn from an open population diagnosed by five experienced Chinese Medical practitioners was therefore performed with all practitioners using the DSOM format. Despite their inexperience with DSOM format, it was found that using the AC2 statistic, the linearly weighted agreement between the practitioners, with agreement by chance removed, was 0.60 ±0.02. The validity of the number is supported by the small standard error, another advantage of the AC2 statistic. This Substantial agreement, as defined by the Landis scale, is a significant improvement on the agreement obtained with the traditional CM diagnostic approach.

Since the practitioners had almost no training and certainly no experience with the DSOM format, the high agreement achieved indicates that the DSOM method of diagnosis could be an appropriate departure point for developing suitable diagnostic format for CM. After a thorough analysis of 60,000 diagnostic records collected over twelve years at UTS Chinese Medical outpatient clinic, it was found that two of the descriptors, Dryness and Phlegm were used the least of the Descriptors. Dryness was never used and Phlegm was the least used, being present in only 0.5% of diagnoses. Wind, while being utilised at the UTS Chinese Medical outpatient clinic, was not among the diagnostic Descriptors in the DSOM.

In agreement with these UTS outpatient clinic data's observations, Dryness and Phlegm were scored above zero the least of the Descriptors by the five practitioners when diagnosing the 42 subjects interviewed. The DSOM format was adjusted to exclude Phlegm and Dryness and include Wind and renamed the Chinese Medicine Diagnostic Descriptor (CMDD) format. The adjustments to the DSOM format enabled any diagnoses in the conventionally used CM format to be readily represented with the CMDD format with the minimum numbers of variables, with no loss of detail.

The CMDD format is proposed as a suitable instrument for describing all contemporary CM diagnoses. The CMDD format comprises of fifteen Descriptors. Scoring each Descriptor from zero to five allows the recording of unlimited CM diagnoses in the one diagnostic form. Agreement is calculated

for non-selection of Descriptors, which is not possible with the CM diagnostic model. The appropriate chance-removed AC2 statistic and standard errors can be estimated with the CMDD format, also not possible with CM diagnoses.

An experiment was implemented with groups of CM practitioners, one group utilising the CMDD and the other the CM diagnostic format diagnosed 35 subjects, three using each format on the first day and two each on the second day. Each of the fifteen CMDD diagnostic Descriptors was scored 0-5, while three selected CM patterns were scored 1-5. The subjects were again drawn from an open population. The level of agreement between the practitioners who used the CMDD was similar to that found between practitioners who used the DSOM format, but significantly larger than the 19% simple agreement mentioned above for practitioners who used the conventional CM format. Indeed, it is shown that when the diagnoses recorded with the conventional CM methodology were concerted to the CMDD format, the simple agreement of 19% dramatically increased to a linearly weighted AC2 inter-rater agreement of 0.67 ±0.03 after removal of agreement by chance. This result is similar to the level of agreement obtained between practitioners who only used the CMDD format, clearly indicating that the low level of agreement when a conventional CM diagnostic is employed is most likely caused by semantic difficulties.

Mapping diagnoses made by raters in the CM to the CMDD format enabled chance-removed inter-rater agreement of 0.65 ±0.03 on day one and

0.73 ±0.03 on day two to be calculated. The agreement calculated in the CMDD format mapped from the CM diagnoses suggests that CMDD seems to better facilitate the diagnostic intention of the rater to be expressed in a format that allows appropriate inter-rate agreement to be calculated than the contemporary CM format.

Moreover, it was shown that when the subjects in both experiments were each divided into three groups on the basis of their overall health status, in both cases the diagnostic agreement between practitioners after removal of agreement by chance was Almost Perfect in the Most Well group. In both cases the chance-removed agreement was Moderate to Substantial in the Least Well group, while the two chance-removed agreements between practitioners for the Intermediately Well cohorts was Slight to Poor.

The DSOM data set was subjected to a number of attempts to use normalisation of data to investigate whether bias was present in the practitioners' scores. These attempts at bias removal did not result in improvements in inter-rater agreement indicating that the practitioners were not biased so that bias was not a significant contributing factor to lowering agreement between them. This observation is contrary to long-held views that a significant proportion of judgement differences between practitioners are due to their differing biases.

Future Work

As part of future work to further improve diagnostic accord between practitioners, the discovery of a lack of practitioner bias suggests that training methods are designed to primarily remove practitioner bias would probably not be effective and that Delphi panel approaches should be explored in the future. Another method could be the development and validation of questionnaires that serve to guide but not replace the practitioners' diagnostic responses in each Descriptor.

The Descriptors need to be re-examined and perhaps redefined, as the use of Liver, Lung or Heart etc. in the CM context could be misleading to patients and/or practitioners, as they are primarily Western medical terms with strong associations in Western culture. On the other hand, the use of these Chinese medical terms alone could also be distracting to practitioners who have not memorized and/or regularly used these terminologies in clinical practice. A conjugation of the Chinese and English terms may be a solution. The use of a standardised Descriptor terminology worldwide would have merit for consistency across languages. This would have similar benefits as the use of Latin terms does for Western medical conditions and anatomy. The development of universal terms for the Descriptors that are independent of languages would be a good outcome for the CM profession. This would have similar benefits as the use of Latin terms does for Western medical conditions and anatomy.

The adjustment and acceptance of the new terms for Descriptors would be a major undertaking. In particular, acceptance of these “new” terms by the profession in the whole world is surely beyond the capacity of any individual and could only be achieved over a long term with the involvement of Professional Organisations in Chinese Medicine in many countries and major International Authorities such as the World Health Organisation.

Subject to further validation and possible adjustment of the CMDD format, investigations of the effectiveness of interventions, as measured by changes in symptoms and confirmed diagnoses in Descriptor(s) values, could commence now. Subjects could be selected by appropriate rating with high scores in the relevant Descriptor(s) by multiple practitioners using the CMDD format. Similarly confirmed diagnoses could be the inclusion criteria for two groups; one clearly categorised as Most Well and the other as Least Well. Objective data of the subjects from each group could be compared to determine if diagnostic markers for definite Descriptor pathology or absence of Descriptor pathology are present. The CMDD appears to be a necessary, vital ingredient required to commence these projects.

The CMDD was developed from the DSOM through a comparison with the UTS outpatient clinic data and an analysis of diagnoses recorded in the DSOM data set made by practitioners trained in the contemporary Chinese Style of CM. The clinic was located in Sydney, which is well known to have a humid climate. As a result the CMDD is specific to the contemporary Chinese Style CM format and quite possibly specific to the climate of the clinic. The

Descriptors identified may be actually different if data from a clinic located in another climatic area had been used and this possibility needs to be investigated. Also, if another style of Traditional East Asian medicine were employed, the consequence might be that a different version of the CMDD may be developed. The most obvious example is Japanese meridian therapy, which due to its apparent simplicity would most likely need fewer Descriptors to adequately describe all diagnostic patterns. It may also be that the Descriptors of the DSOM were indeed appropriate for describing all patterns recorded in Korean style CM.

The findings regarding the poor levels of agreement that occur with the contemporary CM diagnostic format and the vastly superior levels of agreement possible using the CMDD format outlined in this thesis need to published in appropriate journals. This will provide a forum for comment to the CMDD approach. An invitation should next be made to prominent academic figures in Chinese Medicine or perhaps may be delegated by the World Health Organisation, to form a board to facilitate the implementation of the CMDD across the profession.

Analyses of large diagnostic databases from different styles of acupuncture and from different climatic locations would have to be undertaken to determine whether the Descriptors of the CMDD are appropriate to map all diagnoses of any CM style or climatic conditions. Further, experiments will need to be carried out to determine the inter-rater agreement between practitioners in each style.

After this work has taken place, the CMDD in its more completely validated and possibly amended form should be taught in the contemporary Chinese CM curriculum, and introduced to postgraduates as a part of a practitioner's continuing professional education. Future efforts to improve diagnostic reliability after the adoption of the CDDD as the diagnostic format could involve practitioner training and the use of guiding questionnaires.

Due to the levels of diagnostic reliability facilitated with the CMDD, which approach acceptable agreement benchmarks, projects that collect large amounts of clinic data would now be feasible using the CMDD as the diagnostic format. Online patient records from many clinicians could be aggregated and treatments for common conditions examined. The most and least effective interventions could be identified, leading to a 'continuous improvement' effect across the profession. The validation and likely adjustment of the treatments towards more effective health management of patients would improve the outcomes for thousands of patients and enhance the credibility of the Chinese Medicine profession.

Researchers, as a matter of course, should include levels of diagnostic agreement between participating practitioners obtained with the correct statistical tools, as well as the level of inter-rater agreement for each individual Descriptor. Calculations of inter-rater agreement are facilitated by the CMDD format and it therefore should be included as the recommended diagnostic

format in both the internationally accepted STRICTA and CONSORT data collection guidelines.

References

1. Witt, C. *Comparitive Effectiveness Research*. in *SAR - CAAM*. 2014. Beijing: Journal of Integrative Medicne.
2. Witt, C., *Clinical Research on Acupuncture - Concepts and Guidance on Efficacy and Effectiveness Research.* Chinese Journal of Integrative Medicine, 2011. **17**(3): p. 166-172.
3. Witt, C., Aickin, M., Baca, T., Cherkin, D., Haan, M. N., Hammerschlag, R., Hao, J. J., Kaplan, G. A., Lao, L., McKay, T., Pierce, B., Riley, D., Ritenbaugh, C., Thorpe, K., Tunis, S., Weissberg, J., Berman, B. M., *Effectiveness guidance document (EGD) for acupuncture research - a consensus document for conducting trials.* BMC Complementary and Alternative Medicine, 2012. **12**: p. 148.
4. Witt, C., Brinkhaus, B., *Efficacy, effectiveness and cost-effectiveness of acupuncture for allergic rhinitis - An overview about previous and ongoing studies.* Autonomic Neuroscience: Basic and Clinical, 2009. **2010**(157): p. 42-45.
5. Witt, C., Linde, K., *The Need for CAM Research Training.* Forschende Komplementarmedizin, 2008. **15**: p. 69-79.
6. MacPherson, H., Maschino, A. C., Lewith, G., Foster, N. E., Will, C., Vickers, A. J. , *Characteristics of Acupuncture Treatment Associated with Outcome: An Individual Patient Meta-Analysis of 17,922 Patients with Chronic Pain in Randomised Controlled Trials.* PLOS One, 2013. **8**(10): p. e77438.
7. Birch, S.F., R., *Understanding Acupuncture*. 1999, Edinburgh: Churchill Livingstone.
8. Ots, T., Kandirian, A., Szilagyi, S., Sandner-Kielsing, A. *A Meta-Analysis of Sham-Controlled Acupuncture - The Importance of the Dermatomes*. in *2014 International Symposium on Acupuncture Research* 2014. Beijing: Journal of Integrative Medicine.
9. Wenzel, K.W., *Akupuctur: Was zeigen die gerac-Studien? (Acupuncture: What of the GERAC studies?* Deutche Medizinische Wochenschrift, 2005. **130**(24): p. 1520.
10. Haake, M., Muller, H. H., Schade-Brittinger, C., Basler, H. D., Maler, C., Endres, H. G., Trampisch, H. J., Molsberger, A., *German Acupuncture Trials (GERAC) for chronic low back pain: randomized, multicenter, blinded, parrallel-group trail with 3 groups.* Archives of Internal Medicine, 2007. **167**(17): p. 1892-8.
11. Selfe, T.C., Taylor, A. G., *Acupuncture and Osteoarthritis of the Knee.* Fam Community Health, 2008. **31**(3).
12. Cerroni, L., *Lymphoproliferative lesions of the skin.* Journal of Clinical Pathology, 2006. **59**: p. 813-826.
13. Moyer, V.A., Ahn, C., Sneed, S., *Accuracy of Clinical Judgement in Neonatal Jaundice.* Archives of Pediatric Adolescent Medicine, 2011. **154**: p. 391-394.
14. Abdullah, N., Mesurolle, B., El-Koury, M., Kao, E., *Breast Imaging Reporting and Data Systems Lexicon for US: Interobserver Agreement*

for Assessment of Breast Masses. Radiology, 2009. **252**(3): p. 665-672.
15. Joshua, A.M., Celermajer, D. S., Stockler M. R., *Beauty is in the eye of the examiner: Reaching agreement about physical signs and their value.* Internal Medicine Journal, 2005. **35**: p. 178-187.
16. Kahn, C.E., Michalski, T. A., Erickson, S. J., Foley, W. D., Krasnow, A. Z., Lofgren, R. P., Quiroz, F. A., Rand, S. D., *Appropriateness of Imaging Procedure Requests: Do Radiologists Agree?* American Journal of Roentgenology 1996. **169**: p. 11-14.
17. Aboraya, A., Rankin, E., France, C., El-Missiry, A., John, C. , *Reliability of Psychiatric Diagnosis Revisited.* Psychiatry, 2006: p. 41-50.
18. Maciocia, G., *The Foundations of Chinese Medicine*. 1989: Churchill Livingstone.
19. Medicine, S.C.o.T.C., *Acupuncture, a comprehensive Text*. 8th ed. 1981, Seattle: Eastland Press.
20. Macpherson H, T.L., Thomas K, Campbell M., *Acupuncture for lower back pain: Diagnosis and treatment fo 148 patients in a clinical trial.* Complementary Therapies in Medicine, 2004. **12**(1): p. 38-44.
21. Sung, J.J.Y., Leung, W. K., Ching, J. Y. L., Lao, L., Zhang, G., Wu, J. C., Liang, S. M., Xie, H. Ho, Y. P., Chan, L. S., Berman, B., Chan, F. K. L., *Agreements among traditional Chinese medicine practitioners in the diagnosis and treatment of irritable bowel syndrome.* Ailment Parmacol Ther, 2004. **20**: p. 1205-1210.
22. Zhang, G., Bausell, B., Lao, L., Lee, L., Handwerger, B., Berman, B., *The Variability of TCM Pattern Diagnosis and Herbal Prescription on Rheumatoid Arthritis Patients.* Alternative Therapies in Health and Medicine, 2004. **10**(1): p. 58-63.
23. Zhang, G.G., Singh, B., Lee, W., Handwerger, B., Lao, L., Berman, B., *Improvements in Agreement in TCM Diagnosis Among TCM Practitioners for Persons with Conventional Diagnosis of Rheumatoid Arthritis: Effect of Training.* Journal of Alternative and Complementary Medicine, 2008. **14**(4): p. 381-386.
24. O'Brien, K., Birch, S., *A Review of the Reliability of Traditional East Asian Medicine Diagnosis.* The Journal of Alternative and Complimentary Medicine, 2009. **15**(4): p. 353-366.
25. O'Brien, K., Abbas, E., Zhang, J., Guo, Z., Luo, R. Bensoussan, A., Komesaroff, P., *Understanding the Reliability of Diagnostic Variables in a Chinese Medicine Examination.* The Journal of Alternative and Complimentary Medicine, 2008. **15**(7): p. 727-734.
26. O'Brien, K., Abbas, E., Movsessian, P., Hook, M. A., Komesaroff, P., Birch, S. , *Investigating the Reliability of Japanese Toyohari Meridian Therapy Diagnosis.* Journal of Alternative and Complementary Medicine, 2009. **15**(10): p. 1099-1105.
27. Ko, M.M.P., T. Lee, J. A. Choi, T. Kang, B. Lee, M. S., *Interobserver Reliability of Tongue Diagnosis Using Traditional Korean Medicine for Stroke Patients.* Evidence-Based Complimentary and Alternative Medicine, 2011. **2012**(209345).
28. Ko, M.M., Park, T., Lee, J. A., Choi, T., Kang, B., Lee, M. S., *Interobserver Reliability of Pulse Diagnosis Using Traditional Korean*

Medicine for Stroke Patients. Journal of Alternative and Complementary Medicine, 2013. **19**(1): p. 29-34.
29. MacPherson, H., Altman, D.G., Hammersclag, R., Youping, L., Taixaing, W., White, A., Moher, D. , *Revised Standards for Reporting Interventions in Clinical Trails of Acupuncture (Stricta): Extending the CONSORT Statement.* PLoS Med, 2010. **7**(6): p. e1000261.
30. Schulz, K.F., Altman, D. G., Moher, D., *Consort 2010 Statement: updated guidelines for reporting parallel group randomised trials.* British Medical Journal, 2010. **340**: p. 698-702.
31. Zhang, G., 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.* The Journal of Alternative and Complementary Medicine, 2005. **11**(3): p. 415-421.
32. Maciocia, G., *The Practice of Chinese Medicine*. First ed. 1994: Churchill Livingstone.
33. Deng, T., *Practical Diagnosis in Traditional Chinese Medicine*. 1999: Churchill Livingstone.
34. Cambridge. *Cambridge Online Dictionary*. 2015 [cited 2015; Available from: http://dictionary.cambridge.org/dictionary/english/.
35. Cohen, J., *A coefficient of agreement for nominal scales.* Educational and Psychological Measurement, 1960. **xx**(1): p. 37-46.
36. Fleiss, J.L., *Measuring nominal scale agreement among many raters.* Psychological Bulletin, 1971. **76**(5): p. 378-382.
37. Fleiss, J.L., Cohen, J.,, *The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability.* Educational and Psychological Measurement, 1973. **33**: p. 613-619.
38. Brennan, R.L., Prediger, D. L., *Coefficient Kappa: Some uses, misuses, and alternatives.* Educational and Psychological Measurement 1981. **41**: p. 687-699.
39. Byrt, T., Bishop, J., Carlin, J., *Bias, Prevalence and Kappa.* Journal of Clinical Epidemiology, 1993. **46**(5): p. 423-429.
40. Chicchetti, D., Feinstein, A., *High Agreement but Low Kappa: II. Resolving the Paradoxes.* Journal of Epidemiology, 1990. **43**(6): p. 551-558.
41. Cunningham, M., *More than Just the Kappa Coefficient: A Program to Fully Characterize Inter-Rater Reliability between Two Raters.* SAS Global Forum, Statistics and Data Analysis, 2009(242).
42. Feinstein, A., Cicchetti, D., *High Agreement but Low Kappa: I The problems of Two Paradoxes.* Journal of Epidemiology, 1990. **43**(6): p. 543-549.
43. Birkeflet, O.L., P. Vollestad, N., *Low inter-rater reliability in traditional Chinese medicine for female infertility.* Medical Acupuncture, 2011. **29**: p. 51-57.
44. Hua, B., Abbas, E., Hayes, A., Ryan, P., Nelson, L., Obrien, K., *Reliability of Chinese Medicine Diagnosis Variables in the Examination of Patients with Osteoarthritis of the Knee.* Journal of Alternative and Complementary Medicine, 2012. **18**(11): p. 1028-1037.
45. von Eye, A., von Eye, M., *On the Marginal Dependency of Cohen's kappa.* European Psychologist, 2008. **13**(4): p. 305-315.

46. Warrens, M.J., *Inequalities between multi-rater kappas. Advances in Data Analysis and Classification*, 2010: p. Advanced publication doi:10.1007/s11634-010-0073-4.
47. Gwet, K.l., *Kappa Statistic is not Satifactory for Assessing the Extent of Agreement Between Raters.* Statistical Methods For Inter-Rater Reliability Assessment, 2002.
48. Wongpakaran, N., Wongpakaran, T., Wedding, D., Gwet, K., *A Comparison of Cohen's Kappa and Gwet's AC2 when calculating inter-rater reliability coefficients: a study conducted with personality disorder.* Medical Research Methodology, 2013. **2013**(13): p. 61.
49. Randolph, J., *Free-Marginal Multirater Kappa (multirater κfree): An Alternative to Fleiss' Fixed Marginal Multirater Kappa*, in *Learning and Instruction Symposium;* . 2005, University of Joensuu: Joensuu.
50. Gwet, K.l., *Computing inter-rater reliability and its variance in the presence of high agreement.* British Journal of Mathematical and Statistical Psychology, 2008. **61**: p. 29-48.
51. Lo, L.C., Chen, Y. F., Chen, W. J., Cheng, T. L., Chaing, J. Y., *The Study on the Agreement between Automatic Tongue Diagnosis System and Traditional Chinese Medical Practitioners.* Evidence-Based Complimentary and Alternative Medicine, 2012. **2012**(505063).
52. Gwet, K., *AgreeStat 2013.4*. 2013, Advanced Analytics. p. Statistical analysis on the extent of agreement among multiple raters.
53. Landis, J.R., Koch, G. G. , *The measurement of observer agreement for categorical data.* Biometrics, 1977. **33**(1): p. 159-174.
54. Biau, D.J., *Standard Deviation and Standard Error.* Clin Orthop Relat Rs, 2011. **469**: p. 2661-2664.
55. McHugh, M.L., *Interrater reliability: the kappa statistic.* Biochemia Medica, 2012. **22**(3): p. 276-82.
56. Kalinowski, P., *Understanding Confidence Intervals (CIs) and Effect Size Estimatioin.* Observer, 2010. **23**(4).
57. Grant, S.J., Schyner, R. N., Chang, D. HT., Fahey, P., Bensussan, A., *Interrater Reliability of Chinese Medical Diagnosis in People with Prediabetes.* Evidence-Based Complimentary and Alternative Medicine, 2013. **2013**(710892).
58. O'Brien, K., Abbas, E., Zhang, J., Guo, Z., Luo, R., Bensoussan, A., Komesaroff, P., *An Investigation into the Reliability of Chinese Medicine Diagnosis According to Eight Guiding Principles and Zang-Fu Theory in Australians with Hypercholesteralamia.* Journal of Alternative and Complementary Medicine, 2009. **15**(3): p. 259-266.
59. Schnyer, R., Conboy, L., Jacobson, E., McKnight, P., Goddard, T., Moscatelli, F., Legedza, A. Kerr, C., Kaptchuk, T., Wayne, P., *Development of a Chinese Medicine Assessment Measure: An Interdisciplinary Approach Using the Delphi Method.* The Journal of Alternative and Complementary Medicine, 2005. **11**(6): p. 1005-1013.
60. Graham, M., Milanowski, A., Miller, J., *Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings*, C.f.E.C. Reform, Editor. 2012.
61. Barrett, S., *The impact of training on rater variability.* International Education Journal, 2011. **2**(1): p. 49.

62. Pufpaff, L.A., Clarke, L., Jone, R. E., *The Effects of Rater Training on Inter-Rater Agreement.* Mid-Western Educational Researcher 2014. **27**(2).
63. Wang, B., *On Rater Agreement and Rater Training.* English Language Teaching, 2010. **3**(1): p. 108.
64. Joose, P., de Jongh, M, A.C., van Delft-Schreurs, K., Verhofstad, M, H.J., Goslings, J. C., *Improving performance and agreement in injury coding using the Abbreviated Injury Scale: a training course helps.* Health Information Management Journal, 2014. **43**(2): p. 17.
65. Mist, S., Ritenbaugh, C., Aickin, M., *Effects of Questionnaire-Based Diagnosis and Training on Inter-Rater Reliability Among Practitioners of Traditional Chinese Medicine.* The Journal of Alternative and Complementary Medicine, 2009. **15**(7): p. 703-709.
66. Ward, C.H., Beck, A. T., Mendelson, M., Mock, J. E., Erbaugh, J. K., , *The psychiatric Nomenclature Reasons for Diagnostic Disagreement.* Archives of General Psychiatry, 1962. **7**(3): p. 198-205.
67. Cosgrove, L., Drimsky, L. , *A comparison of DSM-IV and DSM-5 panel members' financial associations with industry; A pernicous problem persists.* PLoS Med, 2012. **9**(3): p. 1-5.
68. Greenberg, G., *The Book of Woe. The DSM and the Unmaking of Psychiatry*. 2013, New York: Blue Rider Press. 403.
69. Peele, R., *DSM IV has a label for everyone you might like to treat*, G. Greenberg, Editor. 2011.
70. WHO, *WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region*. 2007, WHO.
71. WHO, *Integrating Traditional Medicine into the WHO Family of International Classifications*. 2010.
72. Alraek, T., Aune, A., Baerheim, A. , *Traditional Chinese medicine syndromes in women with frequently recurring cystitis: frequencies of syndromes and symptoms.* Complementary Therapies in Medicine, 2000. **8**: p. 260-265.
73. Sherman, K., Hogeboom, C., Cherkin, D., *How traditional Chinese medicine acupuncturists would diagnose and treat chronic low back pain: results of a survey of licenced acupunturists in Washington State.* Complementary Therapies in Medicine, 2001. **9**: p. 146-153.
74. Hogeboom, C.J., Sherman, K. J., Cherkin, D. C., *Variation in diagnosis and treatment of chronic low back pain by traditional acupuncturists.* Complementary Therapies in Medicine, 2001. **9**(3): p. 154-166.
75. Zhang, G., Bausell, B., Lao, L., Handwerger, B., Berman, B., *Assessing the consistency of traditional Chinese medical diagnosis: an integrative approach.* Alternative Therapies in Health and Medicine, 2003. **9**(1): p. 66-71.
76. Zhang, G.G., 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.* Journal of Alternative and Complementary Medicine, 2005. **11**(3): p. 415-421.
77. Park, Y., Cho, S., Lee, B., Park, Y. , *Development and Validation of the Yin Deficiency Scale.* Journal of Alternative and Complementary Medicine, 2013. **19**(1): p. 50-56.

78. Park, Y., Park, J., Kim, M., Park, Y. , *The Development of a Valid and Reliable Phlegm Pattern Questionaire.* Journal of Alternative and Complementary Medicine, 2011. **17**(9): p. 851-858.
79. Park, Y., Yang, DH., Lee, JM., Park YB., *Development of a valid and reliable blood stasis questionnaire and its relationship to heart rate variablility.* Complementary Therapies in Medicine, 2013. **21**: p. 633-640.
80. Ryu, H., Lee, H., Kim, H., Kim, J. , *Reliability and Validity of a Cold-Heat Pattern Questionnaire for Traditional Chinese Medicine.* Journal of Alternative and Complementary Medicine, 2010. **16**(6): p. 663-667.
81. Lawrence, D.J., *Chiropractic Management of Low Back Pain and Low Back-Related Leg Complaints: A Literature Synthesis.* Journal of Manipulative and Physiological Therapeutics, 2008. **31**(9): p. 659-674.
82. Muller, U., Roder, C., Greenough, C. G. , *Back related outcome assessment instruments.* Eur Spine J, 2006. **15**(51): p. s25-s31.
83. Jenkinson, C., Coulter, A., Wright, L., *Short Form 36 (SF 36) health survey questionnaire: normative data for adults of working age.* British Medical Journal, 1993. **306**: p. 1437-1440.
84. Gandek, B. *Interpreting the SF-36 Health Survey*. 2002.
85. Taft, C., Karlsson, J., Sullivan, M., *Do SF-36 summary component scores accurately summarize subscale scores?* Quality of Life Research, 2001. **10**(5): p. 395-404.
86. Ware, J., Kosinski, M., *Interpreting SF-36 Summary Health Measures: A Response.* Quality of Life Research, 2001. **10**(5): p. 405-413.
87. Nunaly, J., Bernststein, IH., *Psychometric Theory*. 1994: McGraw-Hill.
88. Newsletter, S.J.T.E.S., *Choosing the Right Type of Rotation in PCA and EFA*, J.D.B.U.o.H.a. Manoa, Editor. 2009.
89. Kim, M.S., Song, K. J., Nam, C. M., Jung, I., *A Study on Comparison of Generalized Kappa Statistics in Agreement Analysis.* The Korean Journal of Applied Statistics, 2012. **25**(5): p. 719-731.
90. Dalkey, N., Helmer, O., *An Experimental Application of the Delphi Method to the Use of Experts.* Management Science, 1963. **9**(3): p. 458-467.
91. Lee, I.O.K., J. W. Ji, G. Y. Lee, Y. T. Kim, G. G. , *Reliability Study for Upgrade of Diagnosis System of Oriental Medicine DSOM(r) S. 1. 1.* Korean Journal of Oriental Physiology & Pathology, 2011. **26**(1).
92. Lee, I.S., Cho, H. S., Um, Y. K., Yu, J. H., Kang, J. G., Kim, K. K., *A Study on Association of DSOM Symptom Scores for Women infertility in Oriental Medicine.* journal of Oriental Obstetrics & Gynecology, 2007. **20**(1): p. 214-238.
93. Lee, I.K., J. Kim, K., *Study on Association of DSOM Items for Uterine Myoma in Oriental Medicine.* J Korean Oriental Medicine, 2007. **28**(2): p. 22-23.
94. Lee, I.S., C. H. Youn, HM. Jung, KK. Kim, K. H. Park, J. Choi, S., *A Study on diagnosis of Dysmenorrhea patients by Diagnosis System of Oriental Medicine (Korean).* journal of Pharmacopuncture, 2007. **10**(1).
95. Lee, I.S., Chi, G. Y., Won, K. J., Lee, Y. T., Kim, K. K. , *Study on Correlation with DSOM Fluents and CBC Biochemical Examination.* Korean Journal of Oriental Physiology & Pathology, 2007. **21**(1): p. 308-317.

96. Cho, H., Kim, JW., Lee, YT., Kim, KK., Lee, IS, *Analysis of Pathogenic Factors in the Menopausal Symptoms of Middle-aged Women in Relation to Sasang Constitutional Type.* Journal of Korean Medicine, 2014. **35**(2): p. 60-68.
97. Transformation, I.f.H.T., *Transforming Health Care Through Big Data*. 2013.
98. Likert, R., *A technique for the measurement of attitudes.* Archives of Psychology, 1932. **22**(140): p. 1-55.
99. O'Brien, K.A., E. Zhang, J. Guo, Z. Luo, R. Bensoussan, A. Komesaroff, P., *An investigation of the reliability of Chinese medicine diagnosis according to eight guiding principles and zang-fu theory in hypercholestorolaemic Australians.* Journal of Alternative and Complementary Medicine, 2009. **15**: p. 259-266.
100. Meir, P.C., *Data from UTS Acupuncture Clinic Database V2*. 2112, UTS.
101. Likert, R., *A Technique for the Measurement of Attitudes.* Archives of Psychology, 1932. **140**: p. 1-55.
102. UTS, *Reliability study of the English version of the Diagnostic System of Oriental Medicine (DSOM) Ethics Approval*, in *Human Research Ethics Committee 2007000152*. 2007.
103. Guangyi, X., Chongsuvivatwong, V., Geater, A., Ming, L., Yun, Z., *Application of Delphi Technique in Identification of Appropriate Screening Questions for Chronic Low Back Pain from Traditional Chinese Medicine Experts' Opinions.* Journal of Alternative and Complementary Medicine, 2009. **15**(1): p. 47-52.
104. Lo, L., Cheng, T. L., Huang, Y. C., Chen, Y. L., Wang, J. T., *Analysis of Agreement on Traditional Chinese Medical Diagnostics for Many Practitioners.* Evidence-Based Complimentary and Alternative Medicine, 2012. **20121**(178081).
105. Quah-Smith, I., Suo, C., Williams, M. A., Sachdev, P. S. , *The Antidepressant Effect of Laser Acupuncture: A Comparison of the Resting Brain's Default Mode Network in Healthy and Depressed Subjects During Functional Magnetic Resonance Imaging.* Medical Acupuncture, 2013. **25**(2): p. 124-133.
106. Spalding, K., Ahn, A., Colbert, A. P., *Acupuncture Needle Stimulation Induces Changes in Bioelectric Potential.* Medical Acupuncture, 2013. **25**(2): p. 141-148.
107. Longhurst, J., *Acupuncture's Cardiovascular Action: A Mechanistic Perspective.* Medical Acupuncture, 2013. **25**(2): p. 101-113.
108. Trochim, W. *The Research Methods Knowledge Base, 2nd Edition* 2006; Available from: <http://www.socialresearchmethods.net/kb/%3E.

Appendices

Appendix 1 The DSOM Questionnaire

DSOM (Diagnosis System of Oriental Medicine)

The aim of this questionnaire is to make a more correct diagnosis through investigating your all ***(sic)*** symptoms. It will take approximately ten minutes to answer all questions. If you are willing to answer these questions sincerely, it will be extremely helpful for us.

This is a multiple-choice test, which will help us to make a diagnosis. Please follow the instructions given and then mark it clearly with a pen so that we can see what you have done.

These questions are all about your health condition. You will be given five possible answers, from ‘of course not’ to ‘exactly’. Please choose the only one answer, which you think is considered to describe you most appropriately by comparing people who are the same age as you.

Facial Symptoms

4. My lips and lower eyelids seem to be bloodless or pale.

4a. I have dark circles under my eyes.

6. My face tends to be flushed or easily changes to a flushed face.

Likes and dislikes in food or taste

20. Normally I should restrain myself from eating certain foods because I am worried I will get indigestion.

20a. If I have some food carelessly, I am likely to have difficulty in digesting food.

10. I have absolutely no idea what is the taste of food, such as boiled rice and bread.

10a. I am suffering from loss of appetite.

The habit of drinking water

33. Recently I often drink cold water, which is due to feeling thirsty and oppressed.

26. It is usual for me to drink water frequently owing to feeling thirsty.

30. I prefer cold water to hot water.

32. I have a bitter taste in my mouth. (When I have some food, It tastes bitter because I feel sick.)

28. I often become parched or very thirsty.

28a. I should make my tongue and lips wet with water or saliva because they often become dry.

Digestive power

17a. I have had difficulty with digesting food since I was young.

17. I cannot digest food well because I often become edgy.

16. I cannot digest food well because I have been edgy lately.

21a. I am often nauseated by something.

21b. After taking western medicines I cannot digest food quickly, and I have a stomachache.

18. After the meal I feel bloated and it takes ages to digest food.

19. I feel tired and drowsy after the meal.

15. I often feel flatulent or have a false sense of satiety.

105. When I feel flatulent, I often have a pain in the abdomen.

22. I often suffer from indigestion.

23. I tend to get carsick frequently.

24. I tend to belch frequently.

Condition of the stool

36. I have constipation and my stool has become hard.

37. When I suffer from constipation, I feel feverish.

50. I often have a dark stool.

A tendency in diarrhea

38. I always go to the toilet before I have a breakfast.

41. Every morning I suffer from diarrhea on rising.

43. I have a soft stool after drinking cold milk or something cold.

44b. I usually do not feel like drinking cold water and cold milk. When I drink them under compulsion, I will definitely have diarrhea.

46. If I am very tense, I often have diarrhea.

46a. If I am very tense, I feel painfully chilly in the stomach and my stool become soft (diarrhea).

40. After evacuation I still feel unsatisfied. (I feel as if there is some stool left which is supposed to come out.)

48a. My stool becomes alternatively soft (diarrhea) and I also suffer from constipation, simply it is not regular.

What is the difference between 48a and 49?

49. I suffer from diarrhea and constipation by turns.

A tendency in your perspiring

51. I usually perspire heavily.

52. Recently I perspire heavily.

53. After perspiring I become exhausted.

54. I break into a cold sweat when sleeping.

55. Even though I hardly move I become tired and sweat.

55b. I often sweat heavily and my skin is cold.

Sensitivity to the heat and cold

57. My body temperature is usually a bit above normal.

57a. I am used to wearing light clothes, because my body temperature is normally a bit high.

58. I do not usually sleep under a blanket.

59. I normally sleep with my feet sticking out of the blanket.

67. Recently my face is glowing with red, and I feel temperature (fever).

60. I am sensitive to the heat but not to the cold.

61. I am sensitive to the cold.

63. My hands and feet are warm.

64. I can feel the heat in my palm.

56. I have high temperature in my hand and feet, besides I feel heavy in the chest.

65a. My hands are quite cold.

65b. My feet are quite cold.

66. If the weather is cold, my hands and feet become cold and look pale (bluish).

106. I have a cold feeling in the abdomen.

141. I feel chilly in the outer opening of the sex organs.

My character

68. I often become angry or fretful.

69. I have hot temper.

69a. When I lose my head or feel excited, I feel the heat in my face.

70. I tend to be capricious so I easily laugh and cry.

71. I am the person who is likely to cry frequently.

72. I am the person who is likely to laugh frequently.

73. I often feel depressed.

74. I often sigh because I feel heavy in the chest or sides.

74a. I feel heavy in the chest due to being worried about something.

75. I often feel nervous.

76. I am likely to be stressed because of my highly sensitive nerves.

79. I have no moment of ease so I am not happy.

A tendency in body pain

88. I feel like lying down because I feel heavy in my body.

90. When it is rainy or cloudy I feel heavier in my body.

90a. I have a haggard face in the morning.

92. After someone gives me a massage, I feel refreshed.

96. I feel chilly and often have a slight general fatigue.

93. When it is rainy, I feel sharp pains over my body.

94. I feel pains depending on how I feel.

95. I often feel sharp pains all over my body.

91. My waist, neck and backbone are stiff and hurt me.

148. In the daytime my symptoms becomes less serious, but at night it becomes more serious.

102. I feel painfully cold in my back.

97. I suffer a pain in my back and waist, and my trunk becomes fatigued.

98. I suffer my cricked back and waist.

98a. I feel heavy in my arms, legs and calves.

99. I feel pain in my waist and knee.

100. I feel exhausted or painfully cold in my waist and knee.

104. I often have a pain in the specific part of the abdomen.

108. I have a stabbing pain in my abdomen.

108a. I often feel pain in the specific part of abdomen, and if I press there with a hand it becomes more serious.

103. My abdomen is hard, and if I press there with a hand it is painful.

109. I feel a pain and stitch in my side.

109a. My pain tends to move around in my body.

109b. If I press between the pit of the stomach and navel it is painful.

When you feel dizzy

84. I often feel dizzy when standing up.

85. I am liable to feel dizzy.

86. I often feel dizzy and have a buzzing in my ears (tinnitus).

87. I feel as if my brain is shaking (headache).

When I feel fatigue

116. It is difficult for me to speak in a loud voice.

116a. After talking a lot with people I feel exhausted.

122. I normally speak in a feeble voice.

123. I have no strength even to breathe.

124. I regard even just chatting as a nuisance because I am exhausted.

119. I am vulnerable to fatigue.

119a. I feel like lying at my ease because I feel tired.

121. I often feel emasculated by something.

125. I have no desire to do anything because I feel tired.

125a. If I work a little hard or sweat a little, my body will be cooled down and I will feel chilly.

About your sleep

130. I cannot get to sleep for a long while at night.

131. I don’t get a sound sleep.

132. I am likely to have a lot of dreams while I am sleeping.

When your skin is dry

134. My hair is lusterless.

135. My fingernails are so weak that they are liable to snap or crack.

136. My heels are often cracked.

147. My lips are often dry and cracked.

137. I have hard skin.

Legs and arms becoming numb

113. While I am sleeping, my arms and legs often go to sleep or become numb.

114. I often have cramps.

115. Sometimes I suddenly have no energy to do anything (I feel as if my hands and feet have no energy to move).

138. I have a haggard skin.

139. I often feel itchy.

Other Questions

83. I cannot put up with being untidy (I am habitually clean and tidy).

127. My fingernails are relatively light colour.

129. My heart sometimes beats violently without any known cause .

129a. I have something wrong with the pit of the stomach and my heart beats violently.

144. I have a slight fever or feel the heat in the afternoon or at night.

145. Phlegm obstructs my throat.

155. I feel severe pain while I am menstruating (menstrual pain).

155a. I have difficult menstruation with clots.

155b. I have a profuse menstruation which seems to be dark.

160. I look pale, and I am often depressed.

165. I often have stiff shoulders without any special reason, and I am bent in the back.

167. I always have a slight cold or have a touch of cold.

168. I often feel shivering with cold.

161. I drink habitually cold water or cold beverages.

161a. I drink plenty of water by nature (unconsciously).

162. After drinking lots of alcoholic beverages, I often have a cough.

164. I often have a dry cough or cough out phlegm.

163. I am often sneezing.

166. If I expose myself to cold weather, I am vulnerable to a fit of sneezing.

Addition

Urine analysis

1. Frequency: How often do you discharge urine a day? (5~6 times a day is normal)

□ Very often □ normal □ relatively rare

2. Quantity: How much do you usually discharge in toilet? (250~300cc for each time is normal)

□ Relatively big amount □ normal □ relatively small amount

3. Feeling: Do you normally finish discharging urine with a feeling of satisfaction?

□ yes I do □ So so □ no I don't (I feel as if I should discharge again)

4. Colour: What colour is your urine?(it is normally light yellow)

□ relatively white like water □ normal □ deep yellow

5. Clearness : How is your urine?

□ it is clear □ normal □ it is thick

Thank you. This is the end of questionnaire.

We appreciate your time and commitment; this information will be a good reference for making a diagnosis in the future.

Appendix 2 Subject Information Statement for participants of the DSOM inter-rater study

UTS
UNIVERSITY
OF TECHNOLOGY
SYDNEY

UNIVERSITY OF TECHNOLOGY, SYDNEY

SUBJECT INFORMATION STATEMENT FOR PARTICIPANTS

Research Project

Title: A reliability study of the English version of the Diagnostic System of Oriental Medicine (DSOM)

(1) What is the study about?

We are conducting an interrater reliability study comparing experienced Chinese Medical practitioners to a Chinese Medical Questionnaire called the DSOM.

(2) Who is carrying out the study?

The study is being carried out by Michael Popplewell, a PhD student under the supervision of Dr Chris Zaslawski from the Department of Medical and Molecular Bioscience, University of Technology. Michael Popplewell can be contacted on 9400 0144 or on his email on michael@wentworthclinic.com. Dr Zaslawski can be contacted on **9516 7856 or on his email Chris.Zaslawski@uts.edu.au** to answer any further questions.

(3) What does the study involve?

You will fill out a questionnaire twice and be diagnosed consecutively by six practitioners.

- **How much time will the study take?**

This study involves approximately two hours of your time at either 10am, 1pm or 3pm on Sunday 29th November 2009.

- **Will I receive any compensation for my time?**

You will receive a payment of a $20 for your participation.

- **Are there any restrictions?**

  You are not permitted to participate in this study if you have a serious illness such as cancer, diabetes or serious heart disease. Colds, muscular injuries or headaches are permitted.

(4) Can I withdraw from the study?

Participating in this study is completely voluntary - you are not under any obligation to consent and there is no problem should you wish to withdraw at any time.

(5) Will the study benefit me?

You will get diagnosed from a Chinese Medical perspective, which may be of interest to you. No treatment will be carried out upon you.

(6) What if there's a problem?

This study has been approved by the University of Technology, Sydney Human Research Ethics Committee. If you have any complaints or reservations about any aspect of your participation in this research which you cannot resolve with the researcher, you may contact the Ethics Committee through the Research Ethics Officer, Ms Susanna Gorman (ph:612 9514 1279). Any complaint you make will be treated in confidence and investigated fully and you will be informed of the outcome.

Appendix 3 DSOM Subject Consent Form

UTS
UNIVERSITY
OF TECHNOLOGY
SYDNEY

UNIVERSITY OF TECHNOLOGY, SYDNEY
CONSENT FORM - STUDENT RESEARCH

Participant

I _____ *(participant's name)* agree to participate in the research project

"**A reliability study of the English version of the**

Diagnostic System of Oriental Medicine (DSOM)".

This project is being conducted by Michael Popplewell as part of his PhD research.

I understand that the purpose of this study is to examine the reliability of the DSOM questionnaire and to compare it to the diagnosis generated by five experienced Chinese Medical (CM) practitioners.

I understand that my participation in this research will involve my being diagnosed by five CM practitioners and filling out a copy of the DSOM twice, which should involve two hours of my time.

I have been provided with a subject information sheet and I am aware that I can contact Michael Popplewell if I have any concerns about the research. I also understand that I am free to withdraw my participation from this research project at any time I wish, without consequences, and without giving a reason.

I agree that Michael Popplewell has answered all my questions fully and clearly.

I agree that the research data gathered from this project may be published in a form that does identify me.

_____ ____/____/____

Signature (participant)

_____ ____/____/____

Signature (researcher or delegate)

NOTE:

This study has been approved by the University of Technology, Sydney Human Research Ethics Committee. If you have any complaints or reservations about any aspect of your participation in this research which you cannot resolve with the researcher, you may contact the Ethics Committee through the Research Ethics Officer (ph: 02 - 9514 9615, Research.Ethics@uts.edu.au), and quote the UTS HREC reference number. Any complaint you make will be treated in confidence and investigated fully and you will be informed of the outcome.

Appendix 4 Diagnostic form for recording the DSOM diagnosis

Eight Questions

Urine

Sleep

Stools

Appetite

Menstruation

Sweating

Fever/Hot Cold

Palpation

Smell

Tongue + Pulse

Pathologies Present

1. Heart

① No Symptom ① Very weak ② Weak ③ Moderate ④ Strong ⑤ Very strong

2. Spleen

① No Symptom ① Very weak ② Weak ③ Moderate ④ Strong ⑤ Very strong

3. Lung

① No Symptom ① Very weak ② Weak ③ Moderate ④ Strong ⑤ Very strong

4. Kidney

① No Symptom ① Very weak ② Weak ③ Moderate ④ Strong ⑤ Very strong

5. Liver

① No Symptom ① Very weak ② Weak ③ Moderate ④ Strong ⑤ Very strong

6. Qi Xu

① No Symptom ① Very weak ② Weak ③ Moderate ④ Strong ⑤ Very strong

7. Yang Xu

① No Symptom ① Very weak ② Weak ③ Moderate ④ Strong ⑤ Very strong

8. Yin Xu ① No Symptom ① Very weak ② Weak ③ Moderate ④ Strong ⑤ Very strong

9. Blood Xu

① No Symptom ① Very weak ② Weak ③ Moderate ④ Strong ⑤ Very strong

10. Damp

① No Symptom ① Very weak ② Weak ③ Moderate ④ Strong ⑤ Very strong

11. Heat

① No Symptom ① Very weak ② Weak ③ Moderate ④ Strong ⑤ Very strong

12. Cold

① No Symptom ① Very weak ② Weak ③ Moderate ④ Strong ⑤ Very strong

13. Dryness

① No Symptom ① Very weak ② Weak ③ Moderate ④ Strong ⑤ Very strong

14. Phlegm

① No Symptom ① Very weak ② Weak ③ Moderate ④ Strong ⑤ Very strong

15. Blood Stagnation

① No Symptom ① Very weak ② Weak ③ Moderate ④ Strong ⑤ Very strong

16. Qi Stagnation

① No Symptom ① Very weak ② Weak ③ Moderate ④ Strong ⑤ Very strong

Appendix 5 Subject Information Statement for Participants of the CMDD study

UNIVERSITY OF TECHNOLOGY SYDNEY

UNIVERSITY OF TECHNOLOGY, SYDNEY

SUBJECT INFORMATION STATEMENT FOR PARTICIPANTS

Research Project

Title: A reliability study comparing Traditional Chinese Medical (TCM) contemporary diagnosis and the Traditional Chinese Medical Diagnostic Descriptor (TCMDD)

(1) What is the study about?

We are conducting an interrater reliability study comparing TCM diagnostic reliability using the normal TCM diagnostic framework and the TCMDD.

(2) Who is carrying out the study?

The study is being carried out by Michael Popplewell, a PhD student under the supervision of Dr Chris Zaslawski from the Department of Medical and Molecular Bioscience, University of Technology. Michael Popplewell can be contacted on 9400 0144 or on his email on michael@wentworthclinic.com. Dr Zaslawski can be contacted on **9516 7856 or on his email Chris.Zaslawski@uts.edu.au** to answer any further questions.

(3) What does the study involve?

You will be diagnosed consecutively by six practitioners.

- **How much time will the study take?**

This study involves approximately two hours of your time at either 10am, 1pm or 3pm on Saturday 8th or Sunday 9th February 2014.

- **Will I receive any compensation for my time?**

You will receive a payment of a $20 for your participation.

(4) Can I withdraw from the study?

Participating in this study is completely voluntary - you are not under any obligation to consent and there is no problem should you wish to withdraw at any time.

(5) Will the study benefit me?

You will get diagnosed from a Chinese Medical perspective, which may be of interest to you. No treatment will be carried out upon you.

(6) What if there's a problem?

This study has been approved by the University of Technology, Sydney Human Research Ethics Committee. If you have any complaints or reservations about any aspect of your participation in this research which you cannot resolve with the researcher, you may contact the Ethics Committee through the Research Ethics Officer, Ms Susanna Gorman (ph:612 9514 1279). Any complaint you make will be treated in confidence and investigated fully and you will be informed of the outcome.

Appendix 6 Consent Form CMDD and CM study

UTS
UNIVERSITY
OF TECHNOLOGY
SYDNEY

UNIVERSITY OF TECHNOLOGY, SYDNEY
CONSENT FORM - STUDENT RESEARCH

Participant

I _____ *(participant's name)* agree to participate in the research project

A reliability study comparing Traditional Chinese Medical (TCM) contemporary diagnosis and the Traditional Chinese Medical Diagnostic Descriptor (TCMDD)

This project is being conducted by Michael Popplewell as part of his PhD research.

I understand that the purpose of this study is to compare a diagnostic system called the TCMDD to the contemporary diagnosis generated by Chinese Medical (CM) practitioners.

I understand that my participation in this research will involve my being diagnosed by six CM practitioners, which should involve two hours of my time.

I have been provided with a subject information sheet and I am aware that I can contact Michael Popplewell on [redacted] or his supervisor Dr Chris Zaslawski on 9514 7856, if I have any concerns about the research. I also understand that I am free to withdraw my

participation from this research project at any time I wish, without consequences, and without giving a reason.

I agree that Michael Popplewell has answered all my questions fully and clearly.

I agree that the research data gathered from this project may be published in a form that does identify me.

_____ ____/____/____

Signature (participant)

_____ ____/____/____

Signature (researcher or delegate)

NOTE:

This study has been approved by the University of Technology, Sydney Human Research Ethics Committee. If you have any complaints or reservations about any aspect of your participation in this research which you cannot resolve with the researcher, you may contact the Ethics Committee through the Research Ethics Officer (ph: 02 - 9514 9615, Research.Ethics@uts.edu.au), and quote the UTS HREC reference number. Any complaint you make will be treated in confidence and investigated fully and you will be informed of the outcome.

Appendix 7 Form for recording the Contemporary Chinese Medical Diagnosis

CM Diagnosis form

Practitioner number …. Patient number…… Interview number……

Three possible diagnoses, each to be valued 1-5

□ Bladder Qi Xu	□ Lung Qi Xu
	□ Lung Yin Xu
□ Blood Xu	□ Phlegm Cold Obstructing the Lung
□ Cold Damp in the Large Intestine	□ Phlegm Heat Disturbing the Heart
□ Cold Damp in the Spleen	□ Phlegm Heat Obstructing the Lung
□ Cold Stagnation in Colon Channel	□ Qi & Blood Stagnation
□ Damp Heat in the Bladder	□ Qi & Blood Stag in the Bladder Channel
□ Damp Heat in the Gall Bladder	□ Qi & Blood Stag in the Gall Bladder Channel
□ Damp Heat in the Large Intestine	□ Qi & Blood stag in the Large Intestine Channel
□ Damp Heat in the Liver	□Qi & Blood Stag in the Small Intestine Channel
□ Damp Heat in the Spleen	□ Qi & Blood Stag in the Spleen Channel
□ Heart and Kidney not Communicating	□ Qi & Blood Stag in the Stomach Channel
□Heart Blood Xu	□ Qi Stagnation (localised trauma)
□ Heart Fire (blazes upwards)	□ Qi Stagnation in the Bladder Channel
□ Heart Qi Xu	□ Qi Stagnation in the Gall Bladder Channel
□ Heart Yin Xu	□ Qi Stagnation in the Large Intestine Channel
□ Heat in the Blood	□ Qi Stagnation in the Liver Channel
□ Heat in the Heart	□ Qi Stagnation in the Small Intestine Channel
□ Heat in the Stomach	□Qi Stagnation in the Stomach Channel
□ Kidney Jing Xu	□ Spleen Blood xu
□ Kidney Qi Xu	□ Spleen Qi Xu

- Kidney Yang Xu
- Kidney Yin Xu
- Liver Blood Xu
- Liver Fire (blazes)
- Liver Heat Rising
- Liver Qi Stagnation
- Liver Wind (Internal - moving)
- Liver Yang Rising
- Spleen Yang Xu
- Stomach Heat
- Stomach Qi xu
- Wind Cold Attacks the Lung
- Wind Heat Attacks the Lung
- Wood (liver) invades Earth (spleen)
- Liver Yin Xu

Appendix 8 Form for recording the CMDD diagnosis

Practitioner number …. Patient number…… Interview number……

CMDD each descriptor valued 0-5

Liver	□	Qi Xu	□	Damp	□
Kidney	□	Yang Xu	□	Wind	□
Lung	□	Yin Xu	□	Heat	□
Spleen	□	Blood Xu	□	Cold	□
Heart	□	Qi Stag	□		
		Blood Stag	□		

Appendix 9a. Score Normalisation by Trip Factor Calculation Table, day one

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Cold	78% -1	-73% +1	10%	-46% +1	51% -1
Heat	17% -1	-18% +1	-45% +1	-24% +1	71% -1
Damp	6%	2%	-53% +1	-5%	56% -1
Dry	27% -1	-60% +1	76% -1	No score	56% -1
Phlegm	63% -1	No score	-9%	18% -1	-9%
Qi Xu	14%	-26% +1	43% -1	-24% +1	-31% +1
Blood Xu	32% -1	77% -1	-41% +1	-19% +1	105% -1
Qi Stag	6%	-4%	22% -1	-32% +1	-12%
Blood Stag	-94% +1	54% -1	10%	-34% +1	84% -1
Yin Xu	27% -1	-60% +1	-56% + 1	-14%	75% -1
Yang Xu	37% -1	-67% +1	53% -1	-27% +1	-4%
Liver	21% -1	-42% +1	16% -1	0%	-4%
Heart	-19% +1	-19% +1	-12%	-35% +1	61% -1
Spleen	-8%	8%	5%	-40% +1	45% -1
Kidney	-28% +1	28% -1	31% -1	6%	31% -1
Lung	14%	31% -1	42% -1	-31% +1	36% -1

Appendix 9b. Score Normalisation by Trip Factor Calculation Table, day two

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Cold	150% -1	-50% +1	0%	No score	0%
Heat	118% -1	-37%+1	-48% +1	4%	-37% +1
Damp	-32% +1	-32% +1	6%	-4%	63% -1
Dry	-26% +1	-15%[5] +1	-26% +1	38% -1	29% -1
Phlegm	5%	-61% +1	-52% +1	92% -1	15% -1
Qi Xu	-26% +1	-14%	-27% +1	38% -1	29% -1
Blood Xu	0%	-55% +1	-77% +1	111% -1	22% -1
Qi Stag	-24% +1	7%	11%	-20% +1	27% -1
Blood Stag	No score	52% -1	63% -1	-78% +1	63% -1
Yin Xu	125% -1	-51% +1	-61% +1	-22% +1	7%
Yang Xu	38% -1	No score	17% -1	91% -1	-46% +1
Liver	-19% +1	-11%	-15[2]% +1	18% -1	26% -1
Heart	44% -1	-74% +1	-24% +1	44% -1	10%
Spleen	-29% +1	-32% +1	-14%	12%	64% -1
Kidney	-3%	-24% +1	-41% +1	26% -1	42% -1
Lung	-10%	-43% +1	-43% +1	56% -1	39% -1

[5] Rounded to one figure after the decimal point but less than 15%

Appendix 9c. Score Normalisation by Trip Factor Calculation Table, day three

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Cold	150% -1	16% -1	No score	-16% +1	-50% +1
Heat	-10%	19% -1	-25% +1	4%	11%
Damp	48% -1	-25% +1	-44% +1	-44% +1	66% -1
Dry	42% -1	-85% +1	28% -1	28% -1	-14%
Phlegm	100% -1	No score	No score	No score	No score
Qi Xu	38% -1	-53% +1	-1%	11%	5%
Blood Xu	27% -1	-65% +1	-18% +1	51% -1	4%
Qi Stag	60% -1	-94% +1	13%	7%	13%
Blood Stag	69% -1	-74% +1	-57% +1	18% -1	44% -1
Yin Xu	-62% +1	-25% +1	64% -1	19% -1	4%
Yang Xu	141% -1	3%	-65% +1	20% -1	No score
Liver	22% -1	-46% +1	17% -1	-25% +1	33% -1
Heart	-6%	6%	-46% +1	60% -1	-13%
Spleen	70% -1	-62% +1	-5%	1%	-5%
Kidney	-19% +1	28% -1	14%	-4%	-19% +1
Lung	250% -1	-75% +1	-75% +1	0%	No score

Appendix 10.a Rater-Descriptor combination scores after Descriptor Normalisation by Trip Factor application, day one

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Average
Cold	12	4	10	8	11	9
Heat	26	36	22	33	41	31.6
Damp	22	23	18	25	24	22.4
Dry	6	5	11	0	8	6
Phlegm	13	0	12	8	12	9
Qi Xu	25	25	28	27	27	26.4
Blood Xu	11	4	11	15	19	12
Qi Stag	36	40	32	34	43	37
Blood Stag	2	19	15	16	21	14.6
Yin Xu	20	12	14	26	28	20
Yang Xu	10	5	10	11	13	9.8
Liver	36	31	34	42	44	37.4
Heart	12	12	14	11	13	12.4
Spleen	40	34	35	30	37	35.2
Kidney	28	26	25	25	24	25.6
Lung	15	16	17	16	32	19.2
Averages	314	292	308	327	397	327.6

Appendix 10.b Rater-Descriptor combination scores after Descriptor Normalisation by Trip Factor application, day two

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Average
Cold	8	4	6	0	6	4.8
Heat	13	10	9	10	10	10.4
Damp	10	9	11	10	12	10.4
Dry	3	0	5	0	0	1.6
Phlegm	2	8	17	5	3	7
Qi Xu	36	29	37	34	32	33.6
Blood Xu	9	5	4	13	8	7.8
Qi Stag	25	27	28	26	21	25.4
Blood Stag	0	9	8	3	11	6.2
Yin Xu	14	8	6	12	11	10.2
Yang Xu	7	0	7	12	5	6.2
Liver	26	22	21	20	22	22.2
Heart	10	4	13	13	13	10.6
Spleen	27	26	23	30	33	27.8
Kidney	23	26	19	22	27	23.4
Lung	11	10	12	12	12	11.4
Averages	224	197	226	222	226	219

Appendix 10.c Rater-Descriptor combination scores after Descriptor Normalisation by Trip Factor application, day three

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Average
Cold	10	4	0	8	5	5.4
Heat	12	10	14	14	15	13
Damp	3	6	6	4	6	5
Dry	7	2	6	6	6	5.4
Phlegm	8	0	0	0	0	1.6
Qi Xu	15	10	15	17	16	14.6
Blood Xu	6	5	10	9	9	7.8
Qi Stag	19	2	19	18	19	15.4
Blood Stag	13	5	9	7	10	8.8
Yin Xu	8	13	15	11	14	12.2
Yang Xu	10	6	3	3	0	4.4
Liver	16	14	15	19	17	16.2
Heart	14	16	13	18	13	14.8
Spleen	19	10	15	16	15	15

Appendix 10.c Rater-Descriptor combination scores after Descriptor Normalisation by Trip Factor application, day three (continued)

Kidney	19	20	24	20	22	21
Lung	9	2	2	4	0	3.4
Averages	188	125	166	174	167	164

Appendix 11a. Differences of Descriptor-rater totals from the mean in Raw and Descriptor Normalised by Trip factor method data, day one

	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5	
Norm Factor	raw	norm	raw	norm	raw	norm	raw	norm	raw	norm
Blood Stag	-15.8	-12.6	9.2	4.4	-1.8	0.4	-5.8	1.4	14.2	6.4
Blood Xu	4.4	-1.0	-10.6	-8.0	-5.6	-1.0	-2.6	3.0	14.4	7.0
Cold	8.8	3.0	-8.2	-5.0	-1.2	1.0	-5.2	-1.0	5.8	2.0
Damp	-1.6	-0.4	-0.6	0.6	-12.6	-4.4	1.4	2.6	13.4	1.6
Dry	2.8	0.0	-6.2	-1.0	7.8	5.0	NS	NS	5.8	2.0
Heart	-2.4	-0.4	-2.4	-0.4	1.6	1.6	-4.4	-1.4	7.6	0.6
Heat	5.8	-5.6	-6.2	4.4	-15.2	-9.6	-8.2	1.4	23.8	9.4
Kidney	-7.6	2.4	-7.6	0.4	8.4	-0.6	-1.6	-0.6	8.4	-1.6
Liver	9.0	-1.4	-18.0	-6.4	7.0	-3.4	0.0	4.6	2.0	6.6
Lung	-2.6	-4.2	-5.6	-3.2	7.4	-2.2	-5.6	-3.2	6.4	12.8
Phlegm	7.0	4.0	NS	NS	1.0	3.0	2.0	-1.0	1.0	3.0
Qi Stag	-2.4	-1.0	1.6	3.0	8.6	-5.0	-12.4	-3.0	4.6	6.0
Qi Xu	-4.2	-1.4	7.8	-1.4	12.8	1.6	-7.2	0.6	-9.2	0.6
Spleen	3.0	4.8	-3.0	-1.2	-2.0	-0.2	-15.0	-5.2	17.0	1.8
Yang Xu	4.6	0.2	-8.4	-4.8	6.6	0.2	-3.4	1.2	0.6	3.2
Yin Xu	6.2	0.0	-13.8	-8.0	-12.8	-6.0	3.2	6.0	17.2	8.0

Appendix 11b. Differences from the mean of Descriptor-rater totals in Raw and Descriptor Normalised by Trip factor method data, day two

	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5	
Norm Factor	raw	norm	raw	norm	raw	norm	raw	norm	raw	norm
Blood Stag	NS	NS	4.8	2.8	5.8	1.8	-7.2	-3.2	5.8	4.8
Blood Xu	0.0	1.2	-5.0	-2.8	-7.0	-3.8	10.0	5.2	2.0	0.2
Cold	9.0	3.2	-3.0	-0.8	0.0	1.2	NS	NS	0.0	1.2
Damp	-3.4	-0.4	-3.4	-1.4	0.6	0.6	-0.4	-0.4	6.6	1.6
Dry	0.6	1.4	-1.4	-1.6	1.6	3.4	-0.4	-1.6	-0.4	-1.6
Heart	5.2	-0.6	-8.8	-6.6	-2.8	2.4	5.2	2.4	1.2	2.4
Heat	11.4	2.6	-3.6	-0.4	-4.6	-1.4	0.4	-0.4	-3.6	-0.4
Kidney	-0.8	-0.4	-5.8	2.6	-9.8	-4.4	6.2	-1.4	10.2	3.6
Liver	-4.6	3.8	-2.6	-0.2	-3.6	-1.2	4.4	-2.2	6.4	-0.2
Lung	-1.2	-0.4	-5.2	-1.4	-5.2	0.6	6.8	0.6	4.8	0.6
Phlegm	-3.6	-5.0	0.4	1.0	5.4	10.0	1.4	-2.0	-3.6	-4.0
Qi Stag	-6.2	-0.4	1.8	1.6	2.8	2.6	-5.2	0.6	6.8	-4.4
Qi Xu	-9.0	2.4	-5.0	-4.6	-9.0	3.4	13.0	0.4	10.0	-1.6
Spleen	-7.8	-0.8	-8.8	-1.8	-3.8	-4.8	3.2	2.2	17.2	5.2
Yang Xu	3.6	0.8	NS	NS	1.6	0.8	8.6	5.8	-4.4	-1.2
Yin Xu	12.8	3.8	-5.2	-2.2	-6.2	-4.2	-2.2	1.8	0.8	0.8

Appendix 11c. Differences from the mean of Descriptor-rater totals in Raw and Descriptor Normalised by Trip factor method data, day three

	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5	
Norm Factor	raw	norm	raw	norm	raw	norm	raw	norm	raw	norm
Blood Stag	9.0	4.2	1.0	-3.8	-6.0	0.2	-1.0	-1.8	-3.0	1.2
Blood Xu	-1.4	-1.8	2.6	-2.8	-3.4	2.2	0.6	1.2	1.6	1.2
Cold	2.6	4.6	-1.4	-1.4	NS	NS	-2.4	2.6	3.6	-0.4
Damp	3.0	-2.0	-6.0	1.0	2.0	1.0	2.0	-1.0	-1.0	1.0
Dry	9.6	1.6	-2.4	-3.4	-2.4	0.6	-2.4	0.6	-2.4	0.6
Heart	5.8	-0.8	-8.2	1.2	-0.2	-1.8	1.8	3.2	0.8	-1.8
Heat	2.4	-1.0	-5.6	-3.0	-1.6	1.0	4.4	1.0	0.4	2.0
Kidney	10.2	-2.0	-15.8	-1.0	2.2	3.0	1.2	-1.0	2.2	1.0
Liver	8.2	-0.2	-8.8	-2.2	-6.8	-1.2	2.2	2.8	5.2	0.8
Lung	-8.4	5.6	-3.4	-1.4	8.6	-1.4	2.6	0.6	NS	NS
Phlegm	8.2	6.4	NS	NS	NS	NS	NS	NS	NS	NS
Qi Stag	4.2	3.6	-8.8	-13.4	3.2	3.6	-4.8	2.6	6.2	3.6
Qi Xu	-1.0	0.4	1.0	-4.6	-7.0	0.4	9.0	2.4	-2.0	1.4
Spleen	11.2	4.0	-9.8	-5.0	-0.8	0.0	0.2	1.0	-0.8	0.0
Yang Xu	-4.0	5.6	6.0	1.6	3.0	-1.4	-1.0	-1.4	-4.0	-4.4
Yin Xu	10.0	-4.2	-3.0	0.8	-3.0	2.8	0.0	-1.2	-4.0	1.8

Appendix 12a. Non-zero scores of each rater-Descriptor combination for calculation of Descriptor Normalisation by Score Factors, day one

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Cold	8	1	5	2	6
Heat	13	9	5	9	16
Damp	9	7	7	8	13
Dry	7	1	7	0	8
Phlegm	5	0	7	5	5
Qi Xu	10	12	14	7	7
Blood Xu	7	1	3	4	9
Qi Stag	16	11	15	9	14
Blood Stag	1	7	7	5	10
Yin Xu	9	3	4	7	12
Yang Xu	7	1	9	3	5
Liver	15	7	15	13	12
Heart	3	3	7	3	7

Appendix 12a. Non-zero scores of each rater-Descriptor combination for calculation of Descriptor Normalisation by Score Factors, day one (continued)

Spleen	14	12	12	8	17
Kidney	9	7	10	6	11
Lung	5	4	8	4	8

Appendix 12b. Non-zero scores of each rater-Descriptor combination for calculation of Descriptor Normalisation by Score Factors, day two

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Cold	7	1	3	0	2
Heat	8	4	4	3	4
Damp	3	2	6	3	5
Dry	1	0	2	1	1
Phlegm	2	2	6	2	1
Qi Xu	11	12	12	13	12
Blood Xu	6	1	2	6	3
Qi Stag	7	10	11	6	11
Blood Stag	0	5	7	1	4
Yin Xu	9	3	2	4	3
Yang Xu	6	0	4	6	1
Liver	7	8	9	9	9
Heart	7	1	4	4	3
Spleen	8	8	10	11	11
Kidney	8	8	5	8	7
Lung	5	3	5	7	5

Appendix 12c Non-zero scores of each rater-Descriptor combination for calculation of Descriptor Normalisation by Score Factors, day three

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Cold	5	3	0	3	2
Heat	5	6	4	5	6
Damp	5	2	3	1	3
Dry	3	1	3	3	3
Phlegm	4	0	0	0	0
Qi Xu	6	3	6	7	7
Blood Xu	5	2	3	4	4
Qi Stag	8	1	7	7	7
Blood Stag	7	2	4	7	7
Yin Xu	3	3	7	5	6

Appendix 12c Non-zero scores of each rater-Descriptor combination for calculation of Descriptor Normalisation by Score Factors, day three (continued)

Yang Xu	4	3	1	4	0
Liver	7	4	7	6	8
Heart	6	5	5	6	5
Spleen	8	4	7	5	6
Kidney	4	7	6	7	5
Lung	5	1	1	2	0

Appendix 13.a Score differences of each rater-Descriptor group from the mean after Descriptor normalisation by Score Factor, day one

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Average
Cold	-8.8	8.2	1.2	5.2	-5.8	11.2
Heat	-5.8	6.2	15.2	8.2	-23.8	33.2
Damp	1.6	0.6	12.6	-1.4	-13.4	23.6
Dry	-2.8	6.2	-7.8	10.2	-5.8	10.2
Phlegm	-7	11	-1	-2	-1	11
Qi Xu	4.2	-7.8	-12.8	7.2	9.2	29.2
Blood Xu	-4.4	10.6	5.6	2.6	-14.4	13.6
Qi Stag	2.4	-1.6	-8.6	12.4	-4.6	38.4
Blood Stag	15.8	-9.2	1.8	5.8	-14.2	16.8
Yin Xu	-6.2	13.8	12.8	-3.2	-17.2	22.8
Yang Xu	-4.6	8.4	-6.6	3.4	-0.6	12.4
Liver	-9	18	-7	0	-2	42
Heart	2.4	2.4	-1.6	4.4	-7.6	12.4
Spleen	-3	3	2	15	-17	37
Kidney	7.6	7.6	-8.4	1.6	-8.4	26.6
Lung	2.6	5.6	-7.4	5.6	-6.4	17.6

Appendix 13.b Score differences of each rater-Descriptor group from the mean after Descriptor normalisation by Score Factor, day two

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Average
Cold	-9	3	0	6	0	6
Heat	-11.4	3.6	4.6	-0.4	3.6	9.6
Damp	3.4	3.4	-0.6	0.4	-6.6	10.4
Dry	-0.6	1.4	-1.6	0.4	0.4	1.4
Phlegm	3.6	-0.4	-5.4	-1.4	3.6	5.6
Qi Xu	9	5	9	-13	-10	34
Blood Xu	0	5	7	-10	-2	9
Qi Stag	6.2	-1.8	-2.8	5.2	-6.8	25.2
Blood Stag	9.2	-4.8	-5.8	7.2	-5.8	9.2
Yin Xu	-12.8	5.2	6.2	2.2	-0.8	10.2
Yang Xu	-3.6	9.4	-1.6	-8.6	4.4	9.4
Liver	4.6	2.6	3.6	-4.4	-6.4	24.6
Heart	-5.2	8.8	2.8	-5.2	-1.2	11.8
Spleen	7.8	8.8	3.8	-3.2	-17.2	26.8
Kidney	0.8	5.8	9.8	-6.2	-10.2	23.8
Lung	1.2	5.2	5.2	-6.8	-4.8	12.2

Appendix 13.c Score differences of each rater-Descriptor group from the mean after Descriptor normalisation by Score Factor, day three

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Average
Cold	-9	-1	6	1	3	6
Heat	1.4	-2.6	3.4	-0.6	-1.6	13.4
Damp	-2.6	1.4	2.4	2.4	-3.6	5.4
Dry	-3	6	-2	-2	1	7
Phlegm	-9.6	2.4	2.4	2.4	2.4	2.4
Qi Xu	-5.8	8.2	0.2	-1.8	-0.8	15.2
Blood Xu	-2.4	5.6	1.6	-4.4	-0.4	8.6
Qi Stag	-10.2	15.8	-2.2	-1.2	-2.2	16.8
Blood Stag	-8.2	8.8	6.8	-2.2	-5.2	11.8
Yin Xu	8.4	3.4	-8.6	-2.6	-0.6	13.4
Yang Xu	-8.2	-0.2	3.8	-1.2	5.8	5.8
Liver	-4.2	8.8	-3.2	4.8	-6.2	18.8
Heart	1	-1	7	-9	2	15

Appendix 13.c Score differences of each rater-Descriptor group from the mean after Descriptor normalisation by Score Factor, day three (continued)

Spleen	-11.2	9.8	0.8	-0.2	0.8	15.8
Kidney	4	-6	-3	1	4	21
Lung	-10	3	3	0	4	4

Appendix 14a Descriptor Normalisation Score Factors raw and rounded and capped, day one

	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5	
Norm Factor	raw	capped	raw	capped	raw	capped	raw	capped	raw	capped
Cold	-1.1	-1	8.2	2	0.2	0	2.6	2	-1.0	-1
Heat	-0.4	-0.5	0.7	0.5	3.0	2	0.9	1	-1.5	-1.5
Damp	0.2	0	0.1	0	1.8	2	-0.2	0	-1.0	-1
Dry	-0.4	-0.5	6.2	2	-1.1	-1	NS		-0.7	-0.5
Phlegm	-1.4	-1.5	NS		-0.1	0	-0.4	-0.5	-0.2	0
Qi Xu	0.4	0.5	-0.7	-0.5	-0.9	-1	1.0	1	1.3	1.5
Blood Xu	-0.6	-0.5	10.6	2	1.9	2	0.7	0.5	-1.6	-1.5
Qi Stag	0.2	0	-0.1	0	-0.6	-0.5	1.4	1.5	-0.3	-0.5
Blood Stag	15.8	2	-1.3	-1.5	0.3	0.5	1.2	1	-1.4	-1.5
Yin Xu	-0.7	-0.5	4.6	2	3.2	2	-0.5	-0.5	-1.4	-1.5
Yang Xu	-0.7	-0.5	8.4	2	-0.7	-0.5	1.1	1	-0.1	0
Liver	-0.6	-0.5	2.6	2	-0.5	-0.5	0.0	0	-0.2	0
Heart	0.8	1	0.8	1	-0.2	0	1.5	1.5	-1.1	-1
Spleen	-0.2	0	0.3	0.5	0.2	0	1.9	2	-1.0	-1
Kidney	0.8	1	1.1	1	-0.8	-1	0.3	0.5	-0.8	-1
Lung	0.5	0.5	1.4	1.5	-0.9	-1	1.4	1.5	-0.8	-1

Appendix 14b Descriptor Normalisation Score Factors raw and rounded and capped, day two

	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5	
Norm Factor	raw	capped	raw	capped	raw	capped	raw	capped	raw	capped
Cold	-1.3	-1.5	3.0	2	0.0	0	NS		0.0	0
Heat	-1.4	-1.5	0.9	1	1.2	1	-0.1	0	0.9	1
Damp	1.1	1	1.7	1.5	-0.1	0	0.1	0	-1.3	-1.5
Dry	-0.6	0.5	NS		-0.8	-1	0.4	-0.5	0.4	-0.5
Phlegm	1.8	0	-0.2	0	-0.9	1	-0.7	-0.5	3.6	2
Qi Xu	0.8	1	0.4	0.5	0.8	1	-1.0	-1	-0.8	-1
Blood Xu	0.0	0	5.0	2	3.5	2	-1.7	-1.5	-0.7	-0.5
Qi Stag	0.9	1	-0.2	0	-0.3	-0.5	0.9	1	-0.6	-0.5
Blood Stag	NS		-1.0	-1	-0.8	-1	7.2	2	-1.5	-1.5
Yin Xu	-1.4	-1.5	1.7	1.5	3.1	2	0.6	0.5	-0.3	-0.5
Yang Xu	-0.6	-0.5	NS		-0.4	-0.5	-1.4	-1.5	4.4	2
Liver	0.7	0.5	0.3	0.5	0.4	0.5	-0.5	-0.5	-0.7	-0.5
Heart	-0.7	-0.5	8.8	2	0.7	0.5	-1.3	-1.5	-0.4	-0.5
Spleen	1.0	1	1.1	1	0.4	0.5	-0.3	-0.5	-1.6	-1.5
Kidney	0.1	0	0.7	0.5	2.0	2	-0.8	-1	-1.5	-1.5
Lung	0.2	0	1.7	1.5	1.0	1	-1.0	-1	-1.0	-1

Appendix 14c. Descriptor Normalisation Score Factors raw and rounded and capped, day three

	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5	
Norm Factor	raw	capped	raw	capped	raw	capped	raw	capped	raw	capped
Cold	-1.8	-2	-0.3	-0.5	NS		0.3	0.5	1.5	1.5
Heat	0.3	0.5	-0.4	-0.5	0.9	1	-0.1	0	-0.3	-0.5
Damp	-0.5	-0.5	0.7	0.5	0.8	1	2.4	2	-1.2	-1
Dry	-1.0	-1	6.0	2	-0.7	-0.5	-0.7	-0.5	0.3	0.5
Phlegm	-2.4	-2	NS		NS		NS		NS	
Qi Xu	-1.0	-1	2.7	2	0.0	0	-0.3	-0.5	-0.1	0
Blood Xu	-0.5	-0.5	2.8	2	0.5	0.5	-1.1	-1	-0.1	0
Qi Stag	-1.3	-1.5	15.8	2	-0.3	-0.5	-0.2	0	-0.3	-0.5
Blood Stag	-1.2	-1	4.4	2	1.7	1.5	-0.3	-0.5	-0.7	-0.5
Yin Xu	2.8	2	1.1	1	-1.2	-1	-0.5	-0.5	-0.1	0
Yang Xu	-2.1	-2	-0.1	0	3.8	2	-0.3	-0.5	NS	

Appendix 14c. Descriptor Normalisation Score Factors raw and rounded and capped, day three (continued)

Liver	-0.6	-0.5	2.2	2	-0.5	-0.5	0.8	1	-0.8	-1
Heart	0.2	0	-0.2	0	1.4	1.5	-1.5	-1.5	0.4	0.5
Spleen	-1.4	-1.5	2.5	2	0.1	0	0.0	0	0.1	0
Kidney	1.0	1	-0.9	-1	-0.5	-0.5	0.1	0	0.8	1
Lung	-2.0	-2	3.0	2	3.0	2	0.0	0	NS	

Appendix 15.a Descriptor Normalising Factors applied in the two methods used day one

	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5	
Norm Factor	TF[6]	SF[7]	TF	SF	TF	SF	TF	SF	TF	SF
Cold	-1	-1	1	2	0	0	1	2	-1	-1
Heat	-1	-0.5	1	0.5	1	2	1	1	-1	-1.5
Damp	0	0	0	0	1	2	0	0	-1	-1
Dry	-1	-0.5	1	2	-1	-1	NS		-1	-0.5
Phlegm	-1	-1.5	NS		0	0	-1	-0.5	0	0
Qi Xu	0	0.5	-1	-0.5	-1	-1	1	1	1	1.5
Blood Xu	-1	-0.5	1	2	1	2	1	0.5	-1	-1.5
Qi Stag	0	0	0	0	-1	-0.5	1	1.5	0	-0.5
Blood Stag	1	2	-1	-1.5	0	0.5	1	1	-1	-1.5
Yin Xu	-1	-0.5	1	2	1	2	0	-0.5	-1	-1.5
Yang Xu	-1	-0.5	1	2	-1	-0.5	1	1	0	0
Liver	-1	-0.5	1	2	-1	-0.5	0	0	0	0
Heart	1	1	1	1	0	0	1	1.5	-1	-1
Spleen	0	0	0	0.5	0	0	1	2	-1	-1
Kidney	1	1	1	1	-1	-1	0	0.5	-1	-1
Lung	0	0.5	1	1.5	-1	-1	1	1.5	-1	-1

[6] TF is an abbreviation for Trip Factor approach
[7] SF is an abbreviation for Score Factor approach

Appendix 15b Descriptor Normalising Factors applied in the two methods used day two

	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5	
Norm Factor	TF	SF	TF	SF	TF	SF	TF	SF	TF	SF
Cold	-1	-1.5	1	2	0	0	NS		0	0
Heat	-1	-1.5	1	1	1	1	0	0	1	1
Damp	1	1	1	1.5	0	0	0	0	-1	-1.5
Dry	1	0.5	NS		-1	-1	-1	-0.5	-1	-0.5
Phlegm	0	0	1	0	1	1	-1	-0.5	1	2
Qi Xu	1	1	0	0.5	1	1	-1	-1	-1	-1
Blood Xu	0	0	1	2	1	2	-1	-1.5	-1	-0.5
Qi Stag	1	1	0	0	0	-0.5	1	1	-1	-0.5
Blood Stag	NS		-1	-1	-1	-1	1	2	-1	-1.5
Yin Xu	-1	-1.5	1	1.5	1	2	1	0.5	0	-0.5
Yang Xu	-1	-0.5	NS		-1	-0.5	-1	-1.5	1	2
Liver	1	0.5	0	0.5	0	0.5	-1	-0.5	-1	-0.5
Heart	-1	-0.5	1	2	1	0.5	-1	-1.5	0	-0.5
Spleen	1	1	1	1	0	0.5	0	-0.5	-1	-1.5
Kidney	0	0	1	0.5	1	2	-1	-1	-1	-1.5
Lung	0	0	1	1.5	1	1	-1	-1	-1	-1

Appendix 15.c Descriptor Normalising Factors applied in the two methods used day three

	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5	
Norm Factor	TF	SF	TF	SF	TF	SF	TF	SF	TF	SF
Cold	-1	-2	-1	-0.5	NS		1	0.5	1	1.5
Heat	0	0.5	-1	-0.5	1	1	0	0	0	-0.5
Damp	-1	-0.5	1	0.5	1	1	1	2	-1	-1
Dry	-1	-1	1	2	-1	-0.5	-1	-0.5	0	0.5
Phlegm	-1	-2	NS		NS		NS		NS	
Qi Xu	-1	-1	1	2	0	0	0	-0.5	0	0
Blood Xu	-1	-0.5	1	2	1	0.5	-1	-1	0	0
Qi Stag	-1	-1.5	1	2	0	-0.5	0	0	0	-0.5
Blood Stag	-1	-1	1	2	1	1.5	-1	-0.5	-1	-0.5
Yin Xu	1	2	1	1	-1	-1	-1	-0.5	0	0
Yang Xu	-1	-2	0	0	1	2	-1	-0.5	1	NS
Liver	-1	-0.5	1	2	-1	-0.5	1	1	-1	-1

Appendix 15.c Descriptor Normalising Factors applied in the two methods used day three (continued)

Heart	0	0	0	0	1	1.5	-1	-1.5	0	0.5
Spleen	-1	-1.5	1	2	0	0	0	0	0	0
Kidney	1	1	-1	-1	0	-0.5	0	0	1	1
Lung	-1	-2	1	2	1	2	0	0	NS	

Appendix 16.a Differences from the mean of Descriptor-rater totals in Raw and Descriptor Normalised by Score Factor data, day one

	Rater1		Rater 2		Rater 3		Rater 4		Rater 5	
Norm Factor	raw	norm	raw	norm	raw	norm	raw	norm	raw	norm
Blood Stag	-15.8	-10.9	9.2	1.6	-1.8	4.6	-5.8	2.1	14.2	2.6
Blood Xu	4.4	2.5	-10.6	-7.0	-5.6	1.0	-2.6	1.0	14.4	2.5
Cold	8.8	2.6	-8.2	-4.4	-1.2	0.6	-5.2	-0.4	5.8	1.6
Damp	-1.6	-1.8	-0.6	-0.8	-12.6	1.2	1.4	1.2	13.4	0.2
Dry	2.8	2.0	-6.2	-2.5	7.8	3.5	NS		5.8	4.5
Heart	-2.4	-0.6	-2.4	-0.6	1.6	1.4	-4.4	-0.6	7.6	0.4
Heat	5.8	1.5	-6.2	0.5	-15.2	-6.0	-8.2	2.0	23.8	2.0
Kidney	-7.6	2.0	-7.6	0.0	8.4	-1.0	-1.6	1.0	8.4	-2.0
Liver	9.0	2.3	-18.0	-6.2	7.0	0.3	0.0	0.8	2.0	2.8
Lung	-2.6	-0.2	-5.6	0.8	7.4	-0.2	-5.6	0.8	6.4	-1.2
Phlegm	7.0	1.5	NS		1.0	3.0	2.0	1.5	1.0	3.0
Qi Stag	-2.4	-1.8	1.6	2.2	8.6	1.7	-12.4	-0.3	4.6	-1.8
Qi Xu	-4.2	1.0	7.8	2.0	12.8	-1.0	-7.2	-2.0	-9.2	0.0
Spleen	3.0	2.4	-3.0	2.4	-2.0	-2.6	-15.0	-1.6	17.0	-0.6
Yang Xu	4.6	2.1	-8.4	-6.4	6.6	3.1	-3.4	-0.4	0.6	1.6
Yin Xu	6.2	4.4	-13.8	-6.1	-12.8	-3.1	3.2	2.4	17.2	2.4

Appendix 16.b Differences from the mean of Descriptor-rater totals in Raw and Descriptor Normalised by Score Factor data, day two

Day 2	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5	
Norm Factor	raw	norm	raw	norm	raw	norm	raw	norm	raw	norm
Blood Stag	NS		52	50	63	33	-78	-33	63	50
Blood Xu	0	14	-55	-37	-77	-24	111	-27	22	20
Cold	150	22	-50	11	0	3	NS		0	33
Damp	-32	0	-32	-5	6	10	-4	0	63	-5
Dry	-26	-36	-15	-100	-26	-9	38	36	29	36
Heart	44	30	-74	-52	-24	6	44	6	10	11
Heat	118	-6	-37	-37	-48	-6	4	4	-37	4
Kidney	-3	0	-24	-4	-41	5	26	-4	42	3
Liver	-19	-8	-11	-9	-15	2	18	-2	26	6
Lung	-10	-10	-43	-5	-43	3	56	3	39	3
Phlegm	5	15	-61	0	-52	-4	92	-4	15	-23
Qi Stag	-24	-2	7	0	11	-11	-20	2	27	4
Qi Xu	-26	3	-14	0	-27	6	38	-2	29	-8
Spleen	-29	2	-32	0	-14	5	12	-8	64	3
Yang Xu	38	49	NS		17	34	91	42	-46	-25
Yin Xu	125	11	-51	0	-61	-16	-22	5	7	0

Appendix 16.c Differences from the mean of Descriptor-rater totals in Raw and Descriptor Normalised by Score Factor data, day three

	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5	
Norm Factor	raw	norm	raw	norm	raw	norm	raw	norm	raw	norm
Blood Stag	9.0	2.0	1.0	-4.0	-6.0	0.0	-1.0	-0.5	-3.0	2.5
Blood Xu	-1.4	0.1	2.6	-1.4	-3.4	0.1	0.6	0.6	1.6	0.6
Cold	2.6	0.4	-1.4	0.9	NS		-2.4	1.9	3.6	1.4
Damp	3.0	0.0	-6.0	-0.5	2.0	0.5	2.0	-0.5	-1.0	0.5
Dry	9.6	0.5	-2.4	-3.5	-2.4	1.0	-2.4	1.0	-2.4	1.0
Heart	5.8	-1.0	-8.2	1.0	-0.2	0.0	1.8	0.0	0.8	0.0
Heat	2.4	1.0	-5.6	-0.5	-1.6	0.5	4.4	0.5	0.4	-1.5
Kidney	10.2	-1.4	-15.8	-0.4	2.2	0.6	1.2	-0.4	2.2	1.6
Liver	8.2	1.3	-8.8	-1.2	-6.8	0.3	2.2	0.8	5.2	-1.2
Lung	-8.4	1.2	-3.4	0.2	8.6	0.2	2.6	1.2	NS	
Phlegm	8.2	2.4	NS		NS		NS		NS	
Qi Stag	4.2	2.0	-8.8	-10.5	3.2	2.0	-4.8	4.5	6.2	2.0

Appendix 16.c Differences from the mean of Descriptor-rater totals in Raw and Descriptor Normalised by Score Factor data, day three (continued)

Qi Xu	-1.0	0.7	1.0	-2.3	-7.0	0.7	9.0	-0.8	-2.0	1.7
Spleen	11.2	0.0	-9.8	-1.0	-0.8	0.0	0.2	1.0	-0.8	0.0
Yang Xu	-4.0	1.8	6.0	1.8	3.0	-0.2	-1.0	0.8	-4.0	-4.2
Yin Xu	10.0	-2.3	-3.0	-0.3	-3.0	1.7	0.0	0.2	-4.0	0.7