Analysis of Herbal Formulation in TCM: Infertility as a case study

Simon K. Poon, Kei Fan, Josiah Poon, Clement Loy, Kelvin Chan
University of Sydney
Sydney, Australia
simon.poon@sydney.edu.au

Xuezhong, Zhou
Beijing Jiaotong University
Beijing, China

Runshun, Zhang, Yinghui
Wang, Jinghong Xie, Baoyan Liu
China Academy of Chinese Medical Sciences
Beijing, China

Paul Kwan
University of New England
Armidale, Australia

Junbin Gao
Charles Sturt University
Australia

Daniel Sze
Hong Kong Polytechnic University, Hong Kong,

Abstract - Herbal prescription in Traditional Chinese Medicine (TCM) relies on the experience of the prescribing doctor after appropriate diagnosis. The combination of herbs chosen, though based on the various Classics, can vary between patients with the same diagnosis. Furthermore, the underlying principles for the choice of herbs are not well defined, and the actual formations can also vary from doctor to doctor. We believe that the prescriptions are the embodiment of the wisdom of the successful doctors, in the process of diagnosis and treatment. Thus it is important to extract the wisdom and discover the knowledge important to the understanding of TCM. The data mining technique employed for analysis here is the Qualitative Comparative Analysis (QCA). This particular analytical method is adopted from social science study, where it is a small-N analysis, meaning that case number is small relative the number of causal factors. The results of this study in the form of configurations indicate not only the herbs (present herbs) necessary for a positive outcome, but the choice of some other herbs (absent herbs) may impact negatively on the outcome.

Keywords - interactions; Causal Complexity, Qualitative Comparative Analysis, Infertility Treatment, Classification Based Association Rules

I. INTRODUCTION

Herbal formation in Traditional Chinese Medicine (TCM) relies on wisdom and expertise of individual practitioner. Although the formations are based on various Classics, the actual formation may vary according to patient’s presentation and practitioner’s interpretation. Successful prescriptions have rarely been captured and analysed and the analytical work would be tedious and difficult. Since the herbal formulae embody the wisdom and expertise of the practitioners, distillation of such knowledge may help to advance the understanding of the underlying principles of the herbal prescription in TCM.

TCM prescriptions depend on not just the herbs that make up a prescription, but the interactions between herbs. The interactions may strengthen the positive effects of a herb, reduce harmful effects, or produce a new effect not seen with only one of the components. Each prescription may contain as many as twenty components selected from a wide range of potential herbs. Quantitative assessment of the effect of prescriptions depends on models that are capable of measuring the complex interactions that are part of the final treatment outcome.

While mass evidence on what constitutes complementarities from interactions has accumulated, the causality of effect is still elusive due to the complex interdependencies of complementarities. The traditional statistical based empirical methods are insufficient for this purpose; due to the fact that these complex systems have relatively fewer cases with large number of factors (variables). The study of Social Science has a similar difficulty. The study employs Qualitative Comparative Analysis (QCA) [1] as a tool for “small-N” analysis.

Successful TCM prescriptions depend on not just the presence of chosen herbs that make up a prescription, but the absence of other herbs may be equally important as the later may impact negatively on the outcome. Discovering intertwining patterns has been a focus of intense research in data mining. Applying data mining technique may distil the configuration where necessary herbs are required for the successful treatment.
The aim of this study is to analyse the complex interdependency among herbs in herbal prescriptions. The framework of the analysis of this study is based on the causality theory adopted from social science study. In this adaptation,

1. Complexity is defined as a bundle viewed as a complex interdependent system exhibiting non-linear dynamics influencing the performance;
2. Configuration is defined as a “fit” among herbs viewed as whole entity or “system”;
3. Confirmation of the discovered configuration is obtained by seeking comments from expert TCM practitioners with the classical knowledge and relevant clinical experiences.

In this study, we apply an efficient method to analyse data such as prescription records for effective configurations of herbs. The output of our framework is thus both a measure of the effectiveness of herbal configurations and the consistencies of the analysis.

II. Method

In this work we apply a two-step framework to analyse causal complexities from TCM patient data record for infertility treatment. This approach integrates two similar techniques to address the concepts of fit. This provides a holistic level analysis of the complex structures of resource interdependencies. It also helps to abstract complexities through the notions of synergistic bundle. The first step of this framework is to extract core herbal components from data using Classification Based Association (CBA) analysis [2]. The second step is to analyse herbal combinations using Qualitative Comparative Analysis (QCA). The aim is to identify herbal combinations that are likely to appear on configuration leading to effective herbal treatment.

![FIGURE I A TWO-STEP FRAMEWORK OF COMPLEXITIES ANALYSIS](image)

Step 1: Identification of Interacting Herbs
Extraction of relevant herbs Classification Based Association (CBA) Analysis

Step 2: Analysis of effective configurations
Analysis of Herbal Combinations using Qualitative Comparative Analysis (QCA)

Results:
Synergistic Configurations of Herbs

A. Classification Based Association (CBA) Analysis

In order to capture the relationships between large numbers of herbal parameters, the data mining technique has been selected. Association analysis (known as Association Rules Mining) has been considered as an appropriate technique in data mining and is perhaps the most common approach to knowledge discovery when there is insufficient evidence of the underlying patterns in a large data set [3]. This technique has proven to be very successful in many fields such as medicine, finance and marketing. In each of these fields, data mining has shed light on various aspects which has assisted in improving the decision making process. This study applies CBA to search for potentially relevant herbs in a given clinical dataset based on two quantitative measures: support and confidence. Association analysis can be viewed as a technique to discover association rules.

An association rule is of the form “if this and this then this”. The left hand side of the rule is called the antecedent or condition and the right hand side is called the consequent or the result. Given a set of item I and a set of transactions T consisting of subset of I, the association rule is of the form X ⇒ Y where X ⊆ I and Y ⊆ I. Here X is called the antecedent and Y is called the consequent. Typically, association rules are considered significant if they satisfy both a minimum support threshold and a minimum confidence threshold. Support (sometimes called coverage) indicates the number of records in the dataset that the rule applies to. In other words, it represents the percentage of transactions that contain both X and Y (i.e. P(X ∪ Y)).

High support means that the rule can be used very often. Confidence indicates how often the rule is correct. It refers to the probability that if the condition is true then the result will also be true (i.e. P(Y | X)). High confidence indicates that the rule is highly dependable.

The rules generated from classification based association analysis could be used in several ways for this analysis. One way is to target the antecedent and get all rules that contain particular items so that it would be possible to identify the items that go together. Another way is to target the consequent so that those items that go with a particular value of the consequent could be identified. Using CBA (a data mining software available for free on the internet [4]), association rules were generated for different values of support and confidence. Finally, it was decided on a set of minimum values required to generate an adequate number of herbs for the qualitative comparative analysis in step 2.

B. Qualitative Comparative Analysis (QCA)

To measure the causality of systems fit among complex interdependent resources, our approach uses the approach of Fuzzy Set Qualitative Comparative Analysis (F5QCA) originally developed by Ragin [5]. This method is part of a family of set-theoretic methods called Qualitative Comparative Analysis. They explain the relationship between the explanatory variables and the outcome variable in the model using sets and subset relations [6]. In general, QCA methods rely on two principle set-subset relationships. They can be used represent the theory for resource or combination of resources being necessary and/or sufficient in causing treatment outcome. A causal condition is
The causal consistency model is given by (1), where $X$ represents the causal set and $Y$ is the outcome set. $m_X$ refers to case $i$’s membership score in the set of the cause $X$, $m_Y$ indicates case $i$’s membership score in the outcome set $Y$ and $m_{X \cap Y}$ defines case $i$’s membership score in the intersection set of $X$ and $Y$.

The steps in fsQCA can be divided into two stages [7]. The first stage involves calibrating the values of the variables for each observation into fuzzy membership scores. The second stage uses the Truth Table algorithm to sort the cases or instances into a truth table in which sufficiency analysis can be carried out [8]. Once the truth table is constructed, it is possible to begin finding configurations in the herbal bundle.

The goal of this step is to aid causal interpretation, in concert with knowledge of the cases. Hence, more generally to explore evidence descriptively and configurationally, with considerations toward the different ways causally relevant conditions may combine to produce given outcome.

### III. ANALYSIS AND RESULTS

Applying the long-term accumulated clinical experiences of TCM in treatment of gynaecological illnesses, recent studies demonstrated that TCM can regulate the gonadotropin-releasing hormone to induce ovulation, and improve uterinal blood flow and menstrual changes of the endometrium. TCM treatment has also been used to treat infertility resulting from polycystic ovarian syndrome, anxiety, stress and immunological disorders [9].

In this study, a dataset of clinical infertility treatment were obtained from the China Academy of Chinese Medical Sciences, Beijing. The dataset captured a total of 250 herbs in combinations in 766 cases (prescriptions) used in the infertility clinics. The dataset were analysed using the 2-step framework described above. After all effective configurations were discovered, the herbs generated were commended for potential clinical significance by expert TCM practitioners.

In this analysis, we have taken a simplified form through bypassing the actual dosages of herbs being used, with the aim to construct a predictor of fertility treatment outcome based on the use or not use of each possible herb in a prescription derived from the Clinical records of treatment. The outcome of each clinical encounter was then classified as good ($1 = $pregnant$)$ or bad ($0 = $not pregnant$)$). From this, the probability of a good outcome was determined for each combination of herbs that could potentially have statistical significance. The summary statistics of the clinical dataset is shown in Table I.

Three of these herbs were frequently used; they can be found in more than 50% of the prescriptions. The most frequently used herb was *tusizi*(菟丝子) with a frequency of 628 times, the second one was *dangguai*(当归), 548 times and *baishao*(白芍) had 467 times.

In the first step, CBA is applied to reduce the numbers of factors (herbs) to identify potential highly interactive herbs that have led to good outcomes. In this process, it is important to keep the number of herbs (often referred as causal conditions in QCA) to be quite low, especially in small-N or intermediate-N research. This is to anticipate the number possible logical combinations of the herbs can quickly exceed the number of cases, and the empirically observed cases will occupy only a small proportion of the potential logical space. This is referred as a “limited diversity” problem [10].

As there is no standard or established method of identifying more appropriate threshold values in CBA, the minimum support and confidence values were chosen by the following rationale. In this step, our focus was to extract herbs that with very high confidence, i.e. herbs or herb sets that are 100% effective in the data set. To do so, we set the support to a relatively low support threshold. The resulting 17 herbs shown in Table II is based on the rules cut-off at support > 1% and confidence = 100%.

### TABLE I BASIC STATISTICS OF THE FERTILITY DATASET

<table>
<thead>
<tr>
<th>Information</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of encounters</td>
<td>766</td>
</tr>
<tr>
<td>Number of herbs</td>
<td>250</td>
</tr>
<tr>
<td>Number of good outcome</td>
<td>404</td>
</tr>
</tbody>
</table>

Based on the results generated from CBA, 17 herbs were selected for subsequent the Qualitative Comparative Analysis. Similar to the previous step, the main goal of this second step was to extract herbal configurations that were high in consistency rather coverage or sample sizes. Unlike research relying in biostatistics, we focused on the scientific validity and utility from the qualitative perspective in small-N research. Note that the common qualitative strategy here was to look for shared causal conditions across multiple instances of the same outcome (e.g. the causally relevant conditions shared by patients that achieve effective treatment). The results are shown in Table III.
TABLE III  CONFIGURATION GENERATED BY QUALITATIVE COMPARATIVE ANALYSIS (QCA) OF 766 CASES CONTAINING THE 17 IDENTIFIED HERBS

| Herb | Person | China Hui | Com | Mate | Qin | Yam | Man | Gao | Jiang | Xi | Wu | 32% | 68% | 13% | Consistency | Coverage |
|------|--------|-----------|-----|------|-----|-----|-----|-----|------|-----|----|-----|-----|-----|-----|-------------|----------|
|      | 1      | 1         | 1   | 1    | 1   | 1   | 1   | 1   | 1    | 1   | 1  | 1   | 0.0425| 1.00 | 0.0245     | 0.98     |
|      | 2      | 2         | 2   | 2    | 2   | 2   | 2   | 2   | 2    | 2   | 2  | 2   | 0.0425| 1.00 | 0.0245     | 0.98     |
|      | 3      | 3         | 3   | 3    | 3   | 3   | 3   | 3   | 3    | 3   | 3  | 3   | 0.0425| 1.00 | 0.0245     | 0.98     |

IV. DISCUSSION AND LIMITATIONS

From the final results, with consistency of 100% and coverage > 0.02 (i.e. 2%), we can extract a herbal triplet (紫苏梗) having a distinctive synergistic effect for infertility treatment. In particular, the couplet (炒白术) and 黄芩 (Scutellaria baicalnsis) is used in a classical formula for the treatment of pregnancy. One of the interesting aspects of the finding is that the most commonly used herbs (such as danggui) amongst the 766 prescriptions were not included in the effective configurations -- either present or absent (Table III). Future investigations may reveal the contribution of the effect of these herbs towards the success of the treatments.

One important limitation of this research was that the confidence threshold in the CBA analysis was too high. Such settings would favor interactions with low support because a 100% confidence is more likely to have arisen by chance in smaller samples sets. This might have led to the more popular herbs with less than perfect effectiveness were removed as candidates to the analysis in step 2. One option was to reduce the confidence threshold (e.g. from 100% to 90%), and to cut down the number of interactions by increasing the support threshold accordingly (e.g. increase to 10%). There could also be a clinical argument that the combinations which represents, for example, less than 10% of the overall dataset, are rare and of lesser clinical interest. Because of the problem addressed was mainly due to limited diversity, we were aware the limitation that the observed data that are far less rich than the potential property space delineated by the conditions.

V. CONCLUSION AND FUTURE WORK

This study demonstrated that our proposed framework can be used to extract synergistic herbal combinations in a variety of clinical situations. By understanding the interdependence among herbs, greater insights on the causality of synergistic impacts can be inferred with the understanding that the question of causality in its own right is a very complex phenomenon.

The significance of this research has two aspects. Firstly, for TCM practitioners - our framework helps to extract evidence based and fine-grained guidelines for treatment decisions. Its purpose also extends to broader drug design. Secondly, for researchers – it enables mid-range theory development of the causality of treatment effectiveness. It proposes novel methods to the analysis of complex interdependencies in herbal prescriptions from the macro viewpoints. Software can also be developed for the research community.

To better understand this issue, we hope to improve our framework by incorporating important perspectives of more recent longitudinal datasets and the ‘dosages’ issue mentioned above. Our research team is currently working on a deeper analysis of the existing datasets from the viewpoint of expert TCM practitioners. Additional datasets are also being prepared for analysis. It is hoped that these analyses will contribute to a more comprehensive understanding of the interactions essential to the effective use and evaluation of TCM treatments; or even help to construct clinically useful TCM recipes in a prospective way to better human life.

ACKNOWLEDGEMENT

This work is partially supported by Programmes of Beijing Municipal S&T Commission, China (D08050703020803 and D08050703020804), China Key Technologies R&D Programme (2007BA110B06-01), and National S&T Major Projects (2009ZX09301-005-010 and 2009ZX10005-019).

REFERENCES


