

Causal Complexities of TCM Prescriptions: Understanding the Underlying Mechanisms of Herbal Formulation

Simon K. Poon, Alan Su, Lily Chau, and Josiah Poon

Abstract Traditional Chinese Medicine (TCM) is a holistic approach to medicine which has been in use in China for thousands of years. The main treatment, Chinese Medicine Formulae is prescribed by combining sets of herbs to address the patient's syndromes and symptoms based on clinical diagnosis. Although herbs are often combined based on various classical formulas, TCM practitioners personalize prescriptions by making adjustments to the formula. However, the underlying principles for the choice of herbs are not well understood. In this chapter, we introduce a framework to explore the complex relationships amongst herbs in TCM clinical prescriptions using Boolean logic. By logically analyzing variations of a large number of TCM herbal prescriptions, we have found that our framework was able to show some herbs may have different pathways to affect effectiveness and such herbs have often been overlooked but can play a weak yet non-trivial role in enhancing the overall effectiveness of the TCM treatment. To achieve this goal, two computational solutions are proposed. An efficient set-theoretic approach is first proposed to study the effectiveness of herbal formulations, and followed by complex network analysis to study the role each herb plays in affecting the outcome.

1 Introduction

TCM prescriptions depend on not just the herbs that make up a prescription, but the inter-relatedness 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 20 components selected from a wide range of potential herbs. Quantitative assessment of the effect of

S.K. Poon (✉) • A. Su • L. Chau • J. Poon
School of Information Technologies, University of Sydney, Sydney, NSW 2006, Australia
e-mail: simon.poon@sydney.edu.au; kasu4088@uni.sydney.edu.au; lily@it.usyd.edu.au;
josiah.poon@sydney.edu.au

prescriptions depends on models that are capable of measuring the complex interactions that are part of the final treatment outcome (Poon et al. 2011a).

Traditionally, analysis of causality has relied on a correlational approach such as multivariate regression, however, it has been demonstrated by several researchers that such an approach cannot account for the phenomena of *conjunctural causation*, *equifinality* and *causal asymmetry* (Ragin 2000; Fiss 2007) which are critically relevant to study causal complexities of TCM.

Conjunctural causation is derived by the fact that an outcome can be achieved from the interaction between multiple causal variables whereas interactions of more than two variables are difficult to interpret using correlational methods such as regression (Fiss 2007). The phenomenon of equifinality suggests that outcomes can be achieved by utilizing different combinations of variables (Katz and Kahn 1978), however, correlational methods such as multivariate regression analysis is unable to account for equifinality as the model produces only a single solution (Fiss 2007). Finally, causal asymmetry addresses the fact that causal relations are asymmetrical in nature (Ehring 1982) which cannot be addressed through correlational analysis as the correlational connections established are symmetrical in nature (Ragin 2008).

Motivated by the inefficiencies with the correlational approach, a new methodology called *Qualitative Comparative Analysis* (QCA) was outlined in (Ragin 1987). QCA is described by Ragin as “an analytic technique designed specifically for the study of cases as configurations of aspects, conceived as combinations of set memberships”. Unlike in correlational methods whereby variables are considered “analytically separable”, the set-theoretic approach combines variables into sets thus enabling its asymmetric nature (Ragin 2008). This is to address the fact that some factors may have asymmetrical effects on outcome. The set of factors that affect positively to effectiveness can be different from the set of factors hindering the effect, i.e. factors that positively affect project success do not necessarily have a reverse effect when they are reduced or removed. In view of the above, we apply an efficient method to analyze data such as prescription records for effective configurations of herbs. The output of our framework is both a measure of the effectiveness of herbal configurations and the consistencies of the analysis.

2 Background of QCA

The implementation of the original QCA that we will discuss here is called *Crisp-Set QCA*, (or *cs/QCA*), which deals with cases that have membership scores that are binary in nature (Ragin 1987). For example, in our application to TCM prescriptions, the membership score for a herb is either zero (0) if it is not used in the prescription, or one (1) if there is a presence of the herb in that particular prescription. Note that the dosage information of herbs is ignored in this study to keep our focus on the logical selection of suitable herbs based on inclusion (or exclusion) of a herb in the TCM prescription. The underpinning procedure in *cs/QCA* is a process of logically eliminating the herbs in the prescription dataset until only the most important herbs remain – this process is termed Boolean minimization.

In cs/QCA, the algorithm used to implement Boolean minimization is called the Quine-McCluskey algorithm first introduced in (Quine 1952, 1955) and later extended in (McCluskey 1956). The algorithm uses a two-level approach similar to solving a *Karnaugh map*.

The first stage of the Quine-McCluskey algorithm generates a set of *prime implicants* from a given truth table. An implicant is defined as a covering of one or more minterms¹ of a Boolean function, and a prime implicant is an implicant that cannot be covered by a more general implicant. The process of generating prime implicants is as follows:

1. Rows in the dataset truth table are grouped based on the number columns with a 1-membership score.
2. Rows in the truth table are combined if they differ by a single variable and this produces an implicant. (e.g. 1,0,1,1,0,1 and 1,1,1,1,0,1 is combined to form the implicant 1,-,1,1,0,1).
3. Repeat step 2 until no more merges are possible in the truth table.
4. Terms which cannot be combined are termed the *prime implicants*.

Once the prime implicants are determined, a prime implicant chart is generated from the output of the first step of the algorithm and the final solution is generated by the second stage of the algorithm. The solution is achieved by removing essential prime implicants, and implicants with row and column dominance and repeating the process until no further reduction can be achieved (Jain et al. 2008).

While the Quine-McCluskey algorithm produces the exact minimal solution for the problem, there is a tradeoff for runtime. It is the problem that is NP-Complete with exponential runtime complexity proportional to the number of causal conditions (Hong et al. 1974; Jain et al. 2008), which presents a major overhead for large-scale analysis. Since the QCA framework was first applied to social and political sciences research, the number of causal conditions that QCA has been used to analyze have been relatively small in quantity and this limitation has gone largely unnoticed. However, in our research, the scale of data analysis that is required is immense as the dataset contains hundreds or even of remedies – for datasets of this magnitude, the Quine-McCluskey algorithm is unable to perform analysis within an adequate timeframe due to the vast number of logical comparisons that will have to be performed.

In order to overcome this issue, we use an alternative algorithm as substitute for the Quine-McCluskey algorithm called BOOM developed by (Fiser and Hlavicka 2003). This algorithm originated from a field of research known as computer aided design and was motivated by the same inefficiencies discussed previously in existing Boolean minimization algorithms. The intended application for the BOOM algorithm was for programmable logic arrays (PLAs), which, similar to our application in TCM prescription data, have vast numbers of variables. Unlike the Quine-McCluskey algorithm which produces an exact solution using a two-level logic

¹ A minterm is a product term of n-variables whereby each variable appears only once. For example, given an input function with variables *a*, *b* and *c*, there are $2^3=8$ minterms, *abc*, *abc'*, *ab'c*, *ab'c'*, *a'bc*, *a'bc'*, *a'b'c*, and *a'b'c'* respectively.

minimization process, the BOOM algorithm produces a near minimal solution using a three-level heuristic approach, which we found to be much more efficient in our testing than previous methodologies.

3 Methodological Implementations

In this work we apply a two-step framework to analyze causal complexities from TCM patient data record for insomnia treatment. This approach integrates two techniques to provide a holistic 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 identify core herbal components from data using Network Analysis (NA). The second step is to analyze herbal prescriptions using a more efficient QCA algorithm. The aim is to identify herbal combinations that are likely to appear on configuration leading to effective herbal treatment, as hidden relationships amongst herbs in prescriptions.

3.1 Network Analysis

A descriptive summary of a binary herb usage data can be visualised with a frequency network. A frequency network can be constructed by drawing an undirected edge for every pair of herbs that is used in one prescription record. The thickness of the edge connecting two herbs increases proportionally to the fraction of the prescription records that contain the herbs together. Where an undirected edge appears in the next set of a prescription record, the edge will increase in weight and thickness. When all edges and weights have been established, the number of edges for each node is computed as a means to adjust the node size. This measure is known as degree centrality in Network Analysis. In a core herb network, a high degree centrality indicates the importance of a herb to working effectively with many other herbs in achieving a treatment. The calculation of the degree centrality is essential in the purpose of breaking ties in next stage of the methodology, the BOOM algorithm.

Strong usage and correlated herbs can be summarised and visualised in a core herb framework. Introduced in the computation of centrality values, a core-herb network based on the frequency of herbs summarises the common herbal combination usage by TCM practitioners. This network is constructed by computing from the raw data, the number of records where Herb A and Herb B are used together. This is the support of the association rule, indicating the proportion of transactions which contain an edge itemset. An undirected edge is therefore constructed between Herb A and Herb B with edge weights determined by the support. For each edge, confidence calculations are also useful in order to determine if the edge itemset has a large percentage of transactions leading to a positive outcome. Confidence in

association rule determines the strength and reliability of the edge itemset. Herb A and Herb B are represented by nodes and the existence of non-zero support and confidence calculations between the two nodes is indicated by an undirected edge. A correlation-based network can similarly be constructed with strength and reliability estimators of herbal combinations.

A core-herb network based on pairwise correlation summarises the association between two herbs as an effective pair. As the raw data is binary, the correlation between herbs is computed using the phi correlation coefficient defined in Eq. 1. The phi coefficient has a maximum value determined by the distribution of the two herb variables A and B. Assuming the data has an equal distribution of positive and negative combinations, the ϕ coefficient will range from -1 to $+1$. ϕ closer to ± 1 indicates strong association while a phi closer to zero indicates weak association.

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\cdot}n_{0\cdot} - n_{\cdot 0}n_{\cdot 1}}} \quad (1)$$

where

$n_{11}, n_{00}, n_{10}, n_{01}$ are record counts of two herb usage; 1 indicates the presence of the herb and 0 indicates the absence of the herb
 n is the total number of observations

Similar to the frequency-based core herb network, Herb A and Herb B is represented by nodes and the correlation between the two nodes is represented by an undirected edge. To estimate the reliability of each correlation coefficient, confidence values are also calculated for each edge. Foundational frequency and correlation networks can therefore be constructed with strength and reliability estimators on each herbal combination edge.

3.2 BOOM Algorithm

The three stages of the BOOM algorithm are *Coverage-Directed Search*, *Implicant Expansion* and *Covering Problem Solution*, respectively. These will now be discussed in detail.

3.2.1 Coverage-Directed Search

The coverage directed search (CD-Search), is named by (Fiser and Hlavicka 2003) as the most innovative part of the algorithm. The algorithm searches for suitable literals (or variables), which are added iteratively to construct an implicant. The strategy for the selection of the initial literal is to use the most frequent literal as it covers the $(n-1)$ -dimensional hypercube. If the $(n-1)$ -dimensional hypercube

found does not intersect with the off-set, it becomes an implicant, otherwise, another literal is added in the same manner described above (Fiser and Hlavicka 2003). One other advantage of the CD-Search is the use of immediate implicant checks when adding literals to a hypercube – when two or more literals have the same frequency, the only ones that will be combined to form a new hypercube is if the new hypercube does not intersect with the off-set. This improves the runtime comparing to the Quine-McCluskey algorithm and generates a higher quality result.

Algorithm 1 CD_Search(F, R) (Fiser and Hlavicka 2003).

Input: F – the set of prescriptions with positive outcomes; R – the set of prescriptions with negative outcomes.

Output: A set of implicants covering F .

```

CD_Search( $F, R$ ) {
   $H = \emptyset$ 
  do
     $F' = F$ 
     $t = \emptyset$ 
    do
       $v = \text{most\_frequent\_literal}(F')$ 
       $t = t \cdot v$ 
       $F' = F' - \text{cubes\_not\_including}(t)$ 
      while ( $t \cap R \neq \emptyset$ )
         $H = H \cup t$ 
       $F = F - F'$ 
    until ( $F == \emptyset$ )
  return  $H$ 
}

```

In the Algorithm (1), F is the on-set, R is the off-set, and H is the set of implicants.

One modification that we have made to the original CD-Search algorithm is that we incorporate the use of domain knowledge in the form of *centrality values* obtained by analyzing the data using network analysis in part A of the methodology. The degree centrality values measure the amount of interaction that a particular herb may have with other herbs in the network. Essentially the centrality values are a measure to influence the selection algorithm when there exists a tie for the most frequent literal – instead of a randomized selection for the most frequent literal, we propose the use of the centrality value ranking as a tie breaker in order to produce a more meaningful result. This approach would favor herbs that have less direct effect on the outcome, but have strong interactions with other herbs, to higher probably to be selected.

Example

Given the data set in Fig. 1, we will follow the BOOM outlined algorithm to find an implicant.

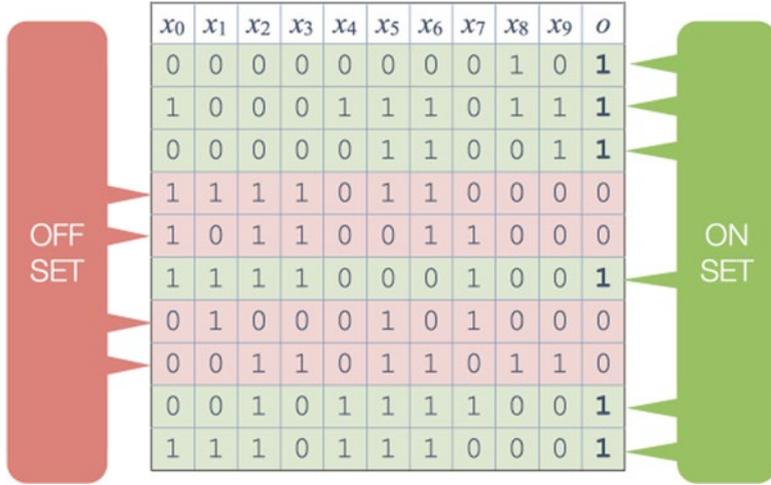


Fig. 1 Example dataset used to demonstrate the BOOM algorithm

In the first iteration, the most common literal in the on-set is x_3' , but as this term intersects with the off-set, it cannot be an implicant, and as a result, another literal will have to be appended (Fig. 2).

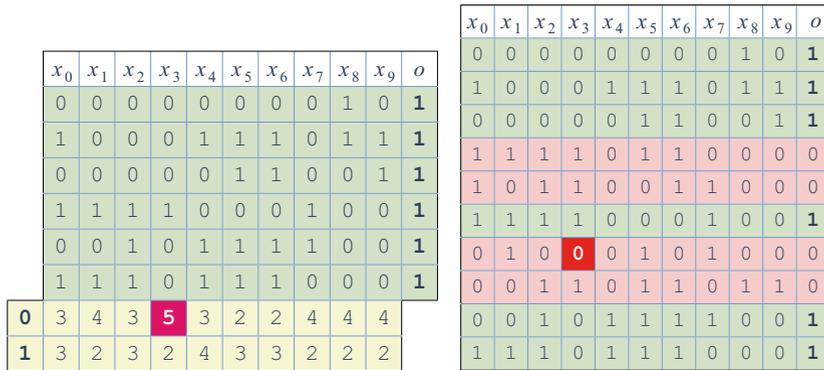


Fig. 2 Coverage-directed search algorithm demonstrating the intersection of x_3 with the off-set

Ignoring the previously discovered term x_3' and the row in the on-set which is not covered by the term, we continue to find the next literal. In this next

step, there are four literals that have the same frequency, in this case, all four combinations with x_3 ' are tried, with the combinations that intersect with the off-set removed and the literal with the greatest centrality value is then chosen from the remaining literals (Fig. 3).

The only combination which intersects with the off-set is x_3 ' x_5 and thus

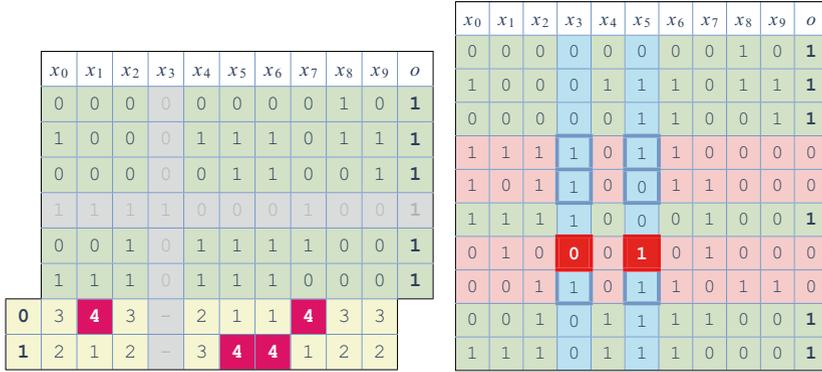


Fig. 3 Coverage-directed search algorithm demonstrating ties in literals

x_3 ' x_1 ', x_3 ' x_6 , and x_3 ' x_7 form the three possible implicant candidates as these sum of products do not intersect with the off-set.

Suppose x_6 has the highest centrality value rank out of the three remaining literals, we choose x_3 ' x_6 as an implicant, and then the next step would be to find another implicant which covers the remainder rows, shown in green in the diagram below (Fig. 4):

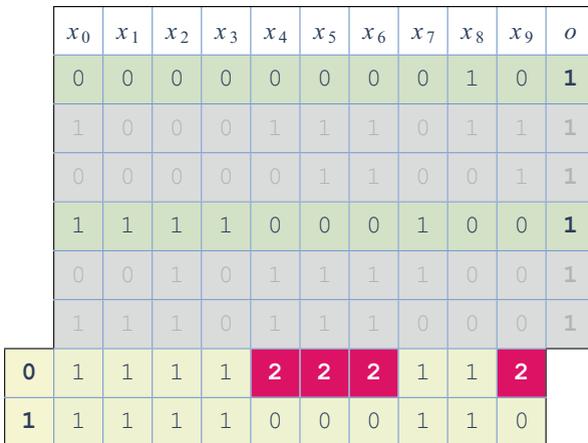


Fig. 4 Coverage-directed search algorithm demonstrating the next stage of finding implicants that cover the remaining rows

The previous steps are repeated until all rows of the on-set are covered and the resultant sum of products (or implicants) is the solution to the coverage-directed search, in this case, a possible solution to the CD-search is $x_3'x_6 + x_5'x_6'$.

The original CD-search algorithm was non-deterministic in nature due to the randomized selection in the presence of multiple literals that are equally frequent. Our modification to the algorithm, which introduces the use of centrality values, aims to eliminate the uncertainty by using a centrality value rank as the tie-break selection criteria.

3.2.2 Implicant Expansion

With the set of implicants generated from the CD-search, the next stage of BOOM called implicant expansion is run in order to produce the prime implicants. The term expansion can be somewhat misleading, as the number of literals in the implicants is actually reduced. However, by removing literals from implicants, their coverage is expanded and thus the name.

Individual literals in each implicant are tried for removal and if the new expression does not intersect with the off-set, then the literal removal is made permanent (Fiser and Hlavicka 2003). There are three strategies for implicant expansion, which are *exhaustive expansion*, *sequential expansion* and *multiple sequential expansion*, respectively. In our testing it was found that the sequential expansion strategy's performance was the most acceptable and the results produced were adequate.

The sequential expansion method simply tries to remove all literals from the implicants one by one and once no further removals are possible, then the newly reduced implicant becomes a prime implicant (Fiser and Hlavicka 2003). One minor downside of this expansion strategy is that it is a greedy algorithm, that is, for each original implicant; only one prime implicant is produced because it does not consider the benefits and costs of removing one implicant as opposed to another. Nonetheless, as noted by the authors, the "simplest sequential expansion is better for very sparse functions" which is the case for our research due to the limited diversity present in the dataset (Fiser and Hlavicka 2003).

3.2.3 Covering Problem Solution

Once the prime implicants are obtained from the implicant expansion process, ideally we would like to reduce the number of prime implicants so that a minimal number of them still cover the given dataset. This is an instance of an NP-hard problem called the *Unate Covering Problem*, i.e. the best known algorithms have exponential

complexity. As noted in (Fiser and Hlavicka 2003), an exact solution to the covering problem is time consuming and that a heuristic approach is the only viable method.

The heuristic proposed in BOOM is called *Least Covered, Most Covering* (LCMC) whereby prime implicants covering minterms which are covered by the least number of other prime implicants are preferred and if there are more than one such prime implicant, then the one which covers the most number of minterms which are not yet covered is chosen (Fiser and Hlavicka 2003).

While the performance of this heuristic is efficient, we felt like the quality of the results could be improved. As a result, we introduce an alternative heuristic as a slight modification to BOOM called *Literal Weights and Output Weights* (WLWO) proposed in (Kagliwal and Balachandran 2012). The Unate Covering Problem can be transformed into a well-known *Set Cover Problem*. This heuristic, unlike the LCMC heuristic, is designed for the sole purpose of logic minimization and takes into account the relationship between implicants and minterms.

This heuristic defines several weights:

1. Literal Weights (LW) – this is defined to be the number of prime implicants which contain such a literal
2. Output Weights (IC) – this is defined to be the number of implicants in the on-set or don't-care-set for each output. In our case with only a single output function, this is simply the cardinality of the on-set and don't-care-set.

Along with the weights, the sub-section then goes on to define several weight functions:

1. Weighted Literal Count (WL): $WL_i = \sum_{x \in X_i} LW_x$
2. Weighted Output Count (WO): $WO_i = \sum_{y \in Y_i} IC_y$

Using these weight functions, the sub-section introduces a three-stage heuristic. Firstly, select prime implicants for inclusion into the final solution if they cover the most number of yet uncovered minterms. If there is a tie, then select the 'shortest' implicant, that is, the one with the lowest literal count. Finally, if there is another tie, then the prime implicant with the highest WLWO heuristic value is used whereby $WLWO_i = WL_i \times WO_i$. (Kagliwal and Balachandran 2012)

The final set of prime implicants produced by the solution to the covering problem forms our final causal configurations with each prime implicant forming a single configuration that leads to the outcome.

3.3 Integration of Results

The set of prime implicants can subsequently be super-imposed on the Herbal Network to verify and determine strong herb-herb interactions and other interesting patterns. Note that as analysing the prime implicant's negative or NOT(herb) result is confounded by the ambiguous definitions of negative, therefore these negative

herbs will be ignored in visualisation. Two comparative core herb networks can therefore be generated and compared to observe interesting patterns; a herb frequency of usage network; and a pairwise herbal correlation network.

Atop either of the two foundational base networks, prime implicants can be super-imposed to visualise interesting results. For each prime implicant set, an undirected edge is created between every pair of herbs in the set. This undirected edge will have a thick line if this edge exists in the base network and a dashed line otherwise. When all edges and weights have been established, the number of edges for each node is computed based on the degrees centrality measures. Interesting factors can thus be inferred by super-imposing positive outcome prime implicants.

4 Data

The described methodology was performed on the insomnia dataset described in (Zhou et al. 2010a). A clinical data warehouse was developed (Zhou et al. 2010b) to integrate and to manage large-scale real-world TCM clinical data. This data warehouse consists of structured electronic medical record from all TCM clinical encounters, including both inpatient and outpatient encounters. There are about 20,000 outpatient encounters of the TCM expert physicians. These encounters included clinical prescriptions for the treatment of various diseases, in which insomnia is a frequently treated disorder.

Total of 460 insomnia outpatient encounters were extracted. The outcome of each encounter was annotated by TCM clinical experts who went through the changes of the insomnia-related variables over consecutive consultation; these include the sleep time per day, sleep quality and difficulty in falling asleep. The outcomes are then classified into two categories: good and bad. When a treatment was effective, which means that if the patient recovered completely or partly from insomnia in the next encounter, then the prescription of the current encounter would be categorized as 'good'; otherwise, the herb prescription would be categorized as 'bad'. After labelling these 460 outpatient encounters, there are 68 encounters with bad outcomes in this dataset; in other words, it is an imbalanced dataset to the advantage of the target class. The average good outcome rate (GOR) of the whole data set is $392/460 = 85.21\%$. There are 261 distinct herbs in the dataset and there are on average 14 herbs in a formula.

5 Results

5.1 Analytical Results from Set-Theoretic

Another important modification to the BOOM algorithm was made such that when two prescriptions are present in the dataset but contributes to both a positive outcome and a negative outcome. Instead of marking these as don't-care terms (whereby

the outcome is marked by ‘-’ instead of 0 or 1), we calculate the ratio of the desirable outcome and the occurrence of this prescription. This is similar to calculating the odds ratios in a case-control study where effectiveness is compared between a set of herbs in a prescription and another prescription with one of more herbs removed. If this ratio were higher than a threshold value, the outcome for this prescription would be set to the desired outcome, otherwise, the undesired outcome. In our case, the threshold used was the overall ratio of desirable outcomes to the total number of prescriptions (Su et al. 2013).

5.1.1 Results from Analysis of Positive Outcomes

We first analyze the causal configurations that lead to a positive outcome, in this case, the on-set of the dataset is set to where the outcome equals to 1. The results produced along with the frequency of these configurations in the prescriptions are shown in Table 1:

5.1.2 Results from Analysis of Negative Results

Next we analyze the causal configurations that attribute to a negative outcome, in this case, the on-set of the dataset is where the outcome is equal to 0. The results produced along with the frequency of these configurations in the prescriptions are as follows (Table 2):

5.2 Results from Network Analysis

Prior to analysis, it is possible to observe descriptive statistics summaries from the described core herbal network. A larger node size indicates a herb is core to a desired outcome. The edge weights between two nodes indicate a dependent association between two nodes. To visualise core herb summaries, both frequency and correlation networks can be generated from the insomnia dataset. The insomnia frequency-based network is shown in Fig. 5 with frequency and confidence of pairwise herb usage indicated on the edges. As the full graph is too dense to quickly extract any important information visually, a threshold of 46 frequency counts was used for visualisation purposes only. This 46 threshold is equivalent to a 10 % support threshold in association rules. The centrality values are derived from the frequency network, as used in the BOOM algorithm. The centrality values are tabulated in Table 3.

After converting the frequencies for each prime implicant (shown in Fig. 5) into pair-wise edge weights, the results generated by this approach are shown to be consistent to the earlier work described in (Zhou et al. 2010b). In regards to the

Table 1 Positive prime implicant results from insomnia dataset

| Configuration | Freq. |
|--|-------|
| ~VAR117•~VAR202•VAR235•VAR237 | 79 |
| ~(陈皮)•~(炒白朮)•(制远志)•(炒酸枣仁) VAR34•~VAR43•VAR200 | 76 |
| (黄连)•~(淡豆豉)•(生甘草) ~VAR43•~VAR120•VAR196•~VAR235 | 53 |
| ~(淡豆豉)•~(五味子)•(大枣)•~(制远志) ~VAR1•~VAR5•~VAR34•~VAR35•~VAR39•~VAR113•~VAR200•~VAR236•~VAR241•~VAR242 | 46 |
| ~(太子参)•~(百合)•~(黄连)•~(黄芩)•~(莲子心)•~(生地黄)•~(生甘草)•~(知母)•~(牡蛎))•~(川芎) VAR175•VAR237 | 45 |
| (山药)•(炒酸枣仁) VAR40•~VAR112•~VAR210•~VAR238 | 44 |
| (石菖蒲)•~(浮小麦)•~(法半夏)•~(柴胡) ~VAR76•VAR117 | 37 |
| ~(竹茹)•(陈皮) ~VAR151•~VAR178•~VAR235•VAR236•VAR237•~VAR242 | 36 |
| ~(煅紫贝齿)•~(炒枳壳)•~(制远志)•(知母)•(炒酸枣仁)•~(川芎) VAR79•VAR203 | 36 |
| (肉桂)•(白芍) VAR33•VAR34•~VAR113•~VAR196•VAR237 | 34 |
| (茯苓)•(黄连)•~(生地黄)•~(大枣)•(炒酸枣仁) ~VAR34•VAR113•VAR235 | 34 |
| ~(黄连)•(生地黄)•(制远志) ~VAR8•~VAR48•~VAR79•~VAR202•~VAR236•VAR237 | 29 |
| ~(党参)•~(夜交藤)•~(肉桂)•~(炒白朮)•~(知母)•(炒酸枣仁) VAR40•~VAR235 | 29 |
| (石菖蒲)•~(制远志) VAR35•VAR237•VAR238 | 28 |
| (黄芩)•(炒酸枣仁)•(柴胡) ~VAR33•~VAR35•~VAR46•~VAR112•~VAR113•~VAR175•~VAR198•~VAR202•~VAR238 | 26 |
| ~(茯苓)•~(黄芩)•~(炒枳实)•~(浮小麦)•~(生地黄)•~(山药)•~(山萸肉)•~(炒白朮)•~(柴胡) ~VAR8•~VAR112•~VAR113•~VAR175•~VAR196•VAR202•~VAR241 | 22 |
| ~(党参)•~(浮小麦)•~(生地黄)•~(山药)•~(大枣)•(炒白朮)•~(牡蛎) VAR117•~VAR202•~VAR210 | 18 |
| (陈皮)•~(炒白朮)•~(法半夏) VAR33•~VAR36•VAR174•VAR236 | 17 |
| (茯苓)•~(麦冬)•(当归)•(知母) VAR43•~VAR196•~VAR238 | 14 |
| (淡豆豉)•~(大枣)•~(柴胡) VAR46•VAR203•~VAR241 | 13 |
| (炒枳实)•(白芍)•~(牡蛎) ~VAR5•VAR35•VAR113•~VAR236 | 13 |
| ~(百合)•(黄芩)•(生地黄)•~(知母) ~VAR34•VAR178•VAR210•VAR238 | |
| ~(黄连)•(炒枳壳)•(法半夏)•(柴胡) | |

Table 2 Negative prime implicant results from insomnia dataset

| Configuration | Freq. |
|---|-------|
| VAR33 •~VAR113•~VAR117•~VAR178•~VAR201•~VAR202•~VAR235•~VAR237 | 14 |
| (茯苓)•~(生地黄)•~(陈皮)•~(炒枳壳)•~(生姜)•~(炒白术)•~(制远志)•~(炒酸枣仁) | |
| ~VAR40•~VAR113•~VAR117•~VAR130• VAR175 •~VAR237 | 11 |
| ~(石菖蒲)•~(生地黄)•~(陈皮)•~(龙齿)•(山药)•~(炒酸枣仁) | |
| VAR33 •~VAR34•~VAR36•~VAR196•~VAR198•~VAR202•~VAR235•~VAR238 | 10 |
| •~VAR241 | |
| (茯苓)•~(黄连)•~(麦冬)•~(大枣)•~(山萸肉)•~(炒白术)•~(制远志)•~(柴胡)•~(牡蛎) | |
| ~VAR34•~VAR40•~VAR84•~VAR113•~VAR117•~VAR196• VAR200 •~VAR202 | 7 |
| •~VAR235 | |
| ~(黄连)•~(石菖蒲)•~(薄荷)•~(生地黄)•~(陈皮)•~(大枣)•(生甘草)•~(炒白术)•~(制远志) | |
| ~VAR79•~VAR113•~VAR117•~VAR175•VAR203•~VAR236• VAR241 | 7 |
| ~(肉桂)•~(生地黄)•~(陈皮)•~(山药)•(白芍)•~(知母)•(牡蛎) | |
| VAR34 •~VAR40•~VAR79•~VAR117•~VAR198•~VAR202•~VAR203•~VAR241 | 7 |
| (黄连)•~(石菖蒲)•~(肉桂)•~(陈皮)•~(山萸肉)•~(炒白术)•~(白芍)•~(牡蛎) | |
| VAR39 •~VAR43•~VAR113•~VAR120•~VAR200•~VAR203•~VAR241•~VAR24 | 7 |
| 2 | |
| (莲子心)•~(淡豆豉)•~(生地黄)•~(五味子)•~(生甘草)•~(白芍)•~(牡蛎)•~(川芎) | |
| VAR76 •~VAR130•~VAR178•~VAR200• VAR210 •~VAR241 | 6 |
| (竹茹)•~(龙齿)•~(炒枳壳)•~(生甘草)•(法半夏)•~(牡蛎) | |
| ~VAR46•~VAR120•~VAR130•~VAR175•~VAR196•~VAR200•~VAR202•~VAR | 6 |
| 235• VAR236 | |
| ~(炒枳实)•~(五味子)•~(龙齿)•~(山药)•~(大枣)•~(生甘草)•~(炒白术)•~(制远志)•(知母) | |
| VAR35 •~VAR40•~VAR46•~VAR79•~VAR113•~VAR210•~VAR237 | 5 |
| (黄芩)•~(石菖蒲)•~(炒枳实)•~(肉桂)•~(生地黄)•~(法半夏)•~(炒酸枣仁) | |
| VAR8 •~VAR35•~VAR79•~VAR174•~VAR200•~VAR201•~VAR242 | 4 |
| (党参)•~(黄芩)•~(肉桂)•~(当归)•~(生甘草)•~(生姜)•~(川芎) | |
| ~VAR35• VAR113 •~VAR174•~VAR202•~VAR237•~VAR242 | 3 |
| ~(黄芩)•(生地黄)•~(当归)•~(炒白术)•~(炒酸枣仁)•~(川芎) | |
| ~VAR34•~VAR35•~VAR39•~VAR48•~VAR200•~VAR201• VAR203 •~VAR210 | 3 |
| ~VAR241•~VAR242 | |
| ~(黄连)•~(黄芩)•~(莲子心)•~(夜交藤)•~(生甘草)•~(生姜)•(白芍)•~(法半夏)•~(牡蛎)•~(川芎) | |
| VAR1 • VAR235 •~VAR238 | 2 |
| (太子参)•(制远志)•~(柴胡) | |
| VAR34 • VAR113 • VAR202 • VAR238 | 1 |
| (黄连)•(生地黄)•(炒白术)•(柴胡) | |

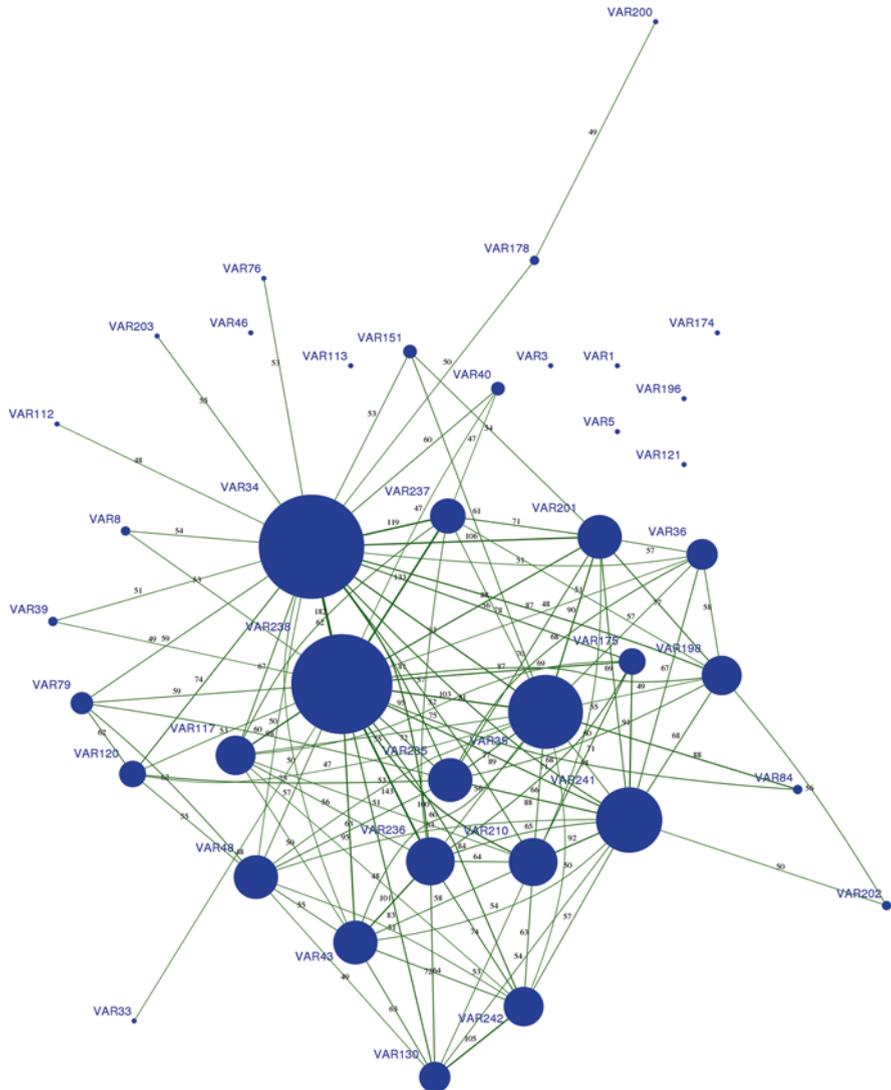


Fig. 5 Energy layout of insomnia frequency-based core herb network with 46 frequency threshold (10 % support threshold). The node size represents frequency of interaction and the edge label highlights the pairwise interaction frequency. (Figure 5 in high resolution with colour can be accessed at <http://www.sydney.edu.au/it/~itcm/book/images/figure2-5.jpg>). The full color version of this image may be viewed in the eBook edition

correlation network, in order to identify reasonably strong pairwise correlations, correlations greater than 0.2 threshold and less than -0.2 threshold of the insomnia dataset is shown in Fig. 6. The ± 0.2 threshold is arbitrarily chosen to avoid a heavily dense network and thus to ensure a quick visual summary of core herbal combinations. Negatively correlated combinations are indicated with a dashed line and

Table 3 Degree centrality values and frequency calculated for each herb, using the full insomnia frequency network

| Herb | Herb name | Chinese name | Centrality | Frequency |
|--------|--|--------------|------------|-----------|
| VAR237 | Stir-frying spine date seed | 炒酸枣仁 | 2,419 | 257 |
| VAR33 | Indian bread | 茯苓 | 2,249 | 253 |
| VAR34 | Golden thread | 黄连 | 1,580 | 165 |
| VAR238 | Chinese thorowax root | 柴胡 | 1,576 | 174 |
| VAR235 | Prepared thinleaf milkwort root | 制远志 | 1,554 | 166 |
| VAR200 | Fresh liquorice root | 生甘草 | 1,456 | 165 |
| VAR203 | White peony root | 白芍 | 1,377 | 153 |
| VAR236 | Common anemarrhena rhizome | 知母 | 1,362 | 151 |
| VAR210 | Prepared pinellia tuber | 法半夏 | 1,358 | 148 |
| VAR196 | Chinese date | 大枣 | 1,248 | 132 |
| VAR174 | Chinese angelica | 当归 | 1,200 | 139 |
| VAR40 | Grassleaf sweetflag rhizome | 石菖蒲 | 1,167 | 130 |
| VAR113 | Dried/fresh rehmannia [root] | 生地黄 | 1,127 | 131 |
| VAR241 | Oyster shell | 牡蛎 | 1,111 | 120 |
| VAR121 | Dragon bone | 龙骨 | 1,028 | 105 |
| VAR79 | Cassia bark | 肉桂 | 995 | 98 |
| VAR117 | Dried tangerine peel | 陈皮 | 972 | 106 |
| VAR46 | Stir-frying immature orange fruit | 炒枳实 | 969 | 106 |
| VAR242 | Szechwan lovage rhizome | 川芎 | 958 | 112 |
| VAR35 | Baical skullcap root | 黄芩 | 940 | 109 |
| VAR76 | Bamboo shavings | 竹茹 | 858 | 87 |
| VAR39 | Lotus plumule | 莲子心 | 754 | 79 |
| VAR130 | Dragon teeth | 龙齿 | 750 | 77 |
| VAR202 | Stir-frying largehead atractylodes rhizome | 炒白术 | 673 | 86 |
| VAR5 | Lily bulb | 百合 | 640 | 72 |
| VAR36 | Dwarf lilyturf tuber | 麦冬 | 617 | 84 |
| VAR48 | Tuber fleecflower stem | 夜交藤 | 594 | 74 |
| VAR8 | Tangshen | 党参 | 565 | 64 |
| VAR112 | Light wheat | 浮小麦 | 565 | 58 |
| VAR201 | Fresh ginger | 生姜 | 557 | 58 |
| VAR175 | Common yam rhizome | 山药 | 536 | 70 |
| VAR84 | Peppermint | 薄荷 | 510 | 54 |
| VAR198 | Asiatic cornelian cherry fruit | 山萸肉 | 455 | 58 |
| VAR120 | Chinese magnoliavine fruit | 五味子 | 372 | 46 |
| VAR178 | Stir-frying orange fruit | 炒枳壳 | 357 | 50 |
| VAR3 | Phyllanthus ussuriensis | 蜜甘草 | 317 | 40 |
| VAR43 | Fermented soybean | 淡豆豉 | 278 | 33 |
| VAR151 | Arabic cowry shell | 煅紫贝齿 | 275 | 26 |
| VAR1 | Heterophylly falsestarwort root | 太子参 | 121 | 14 |

positively correlated combinations with a solid line where phi correlation values are indicated on the edges. Foundational frequency and correlation networks can therefore be constructed with strength and reliability estimators on each herbal combination edge.

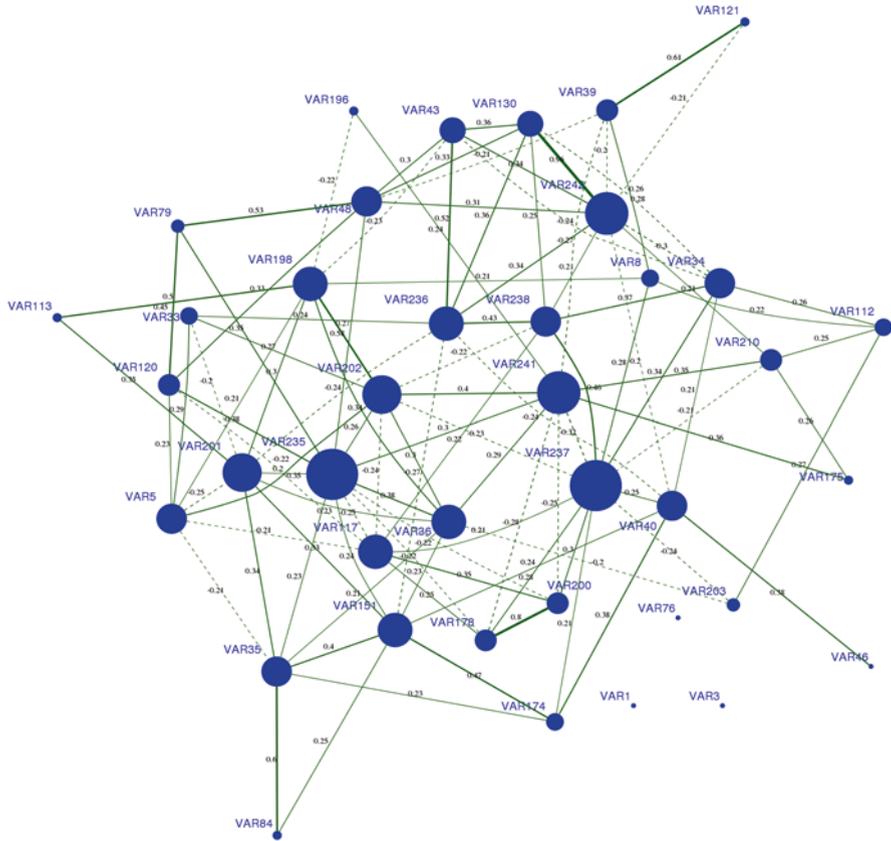


Fig. 6 Energy layout of insomnia correlation-based core herb network with ± 0.2 threshold and correlation weighted edge weights, *dotted edges* indicate negative correlation, and node size indicate frequency of interaction for the given node. (Figure 6 in high resolution with colour can be accessed at <http://www.sydney.edu.au/it/~itcm/book/images/figure2-6.jpg>)

5.3 Results from Interaction Analysis

Frequency and correlation networks can suggest possible interactions between herbs that are interesting to test for. As the reduced frequency network displayed the majority of threshold prime implicant sets, only the cycles will be analysed as possible pair-wise interactions. Mathematically, this is expressed in Eq. 2 and can be similarly expanded for higher dimensions (e.g. Poon et al. 2011b).

$$(n_{11} + n_{00}) > (n_{01} + n_{10}) \tag{2}$$

where n_{ij} is the frequency of Herb i and Herb j where 1 indicates the presence of a herb and 0, absence.

Table 4 Interaction analysis of pairwise combinations derived from prime implicant connections in frequency and correlation networks

| | Combination | Frequency | Good outcomes | Good outcome rate | Additiveness |
|--------|-------------|-----------|---------------|-------------------|----------------|
| VAR241 | 11+00 | 35+122 | 31+102 | 84.71 | Sub-additive |
| VAR33 | 01+10 | 218+85 | 190+69 | 85.48 | |
| VAR210 | 11+00 | 81+140 | 70+112 | 82.35 | Sub-additive |
| VAR33 | 01+10 | 172+67 | 151+59 | 87.87 | |
| VAR210 | 11+00 | 46+238 | 41+204 | 86.27 | Super-additive |
| VAR241 | 01+10 | 74+102 | 59+88 | 83.52 | |
| VAR175 | 11+00 | 46+179 | 46+142 | 83.56 | Sub-additive |
| VAR237 | 01+10 | 211+24 | 191+13 | 86.81 | |
| VAR113 | 11+00 | 56+219 | 53+174 | 82.55 | Super-additive |
| VAR235 | 01+10 | 110+75 | 99+66 | 89.19 | |
| VAR113 | 11+00 | 51+227 | 47+186 | 83.81 | Super-additive |
| VAR203 | 01+10 | 102+80 | 87+72 | 87.36 | |
| VAR113 | 11+00 | 24+244 | 21+200 | 82.46 | Super-additive |
| VAR35 | 01+10 | 85+107 | 73+98 | 89.06 | |
| VAR203 | 11+00 | 34+232 | 31+195 | 84.96 | Super-additive |
| VAR35 | 01+10 | 75+119 | 63+103 | 85.57 | |
| VAR200 | 11+00 | 90+220 | 89+180 | 86.77 | Super-additive |
| VAR34 | 01+10 | 75+75 | 60+63 | 82.00 | |
| VAR237 | 11+00 | 48+142 | 47+108 | 81.58 | Sub-additive |
| VAR35 | 01+10 | 61+209 | 47+190 | 87.78 | |
| VAR235 | 11+00 | 143+180 | 132+135 | 82.66 | Super-additive |
| VAR237 | 01+10 | 114+23 | 105+20 | 91.24 | |
| VAR203 | 11+00 | 46+247 | 43+214 | 87.71 | Super-additive |
| VAR46 | 01+10 | 60+107 | 44+91 | 80.84 | |
| VAR113 | 11+00 | 34+293 | 34+248 | 86.24 | Super-additive |
| VAR175 | 01+10 | 36+97 | 25+85 | 82.71 | |
| VAR203 | 11+00 | 100+150 | 93+114 | 82.80 | Super-additive |
| VAR237 | 01+10 | 157+53 | 144+41 | 88.10 | |
| VAR238 | 11+00 | 37+211 | 34+169 | 81.85 | Super-additive |
| VAR242 | 01+10 | 75+137 | 68+121 | 89.15 | |
| VAR33 | 11+00 | 32+175 | 32+144 | 85.02 | Sub-additive |
| VAR8 | 01+10 | 32+221 | 27+189 | 85.38 | |

Table 4 summarises results of pairwise interaction analysis. Interpretation of the interaction analysis results will not be described here as the purpose of this chapter is to introduce a cause and effect methodology between herbs. Common node and cycle analysis can therefore identify effective higher order combinations by taking advantage of low correlated or low frequently ignored herbs.

5.4 Overall Results and Visualization

In order to visualise causal herb combinations, prime implicants are superimposed on the base frequency and correlation networks. The resulting frequency and

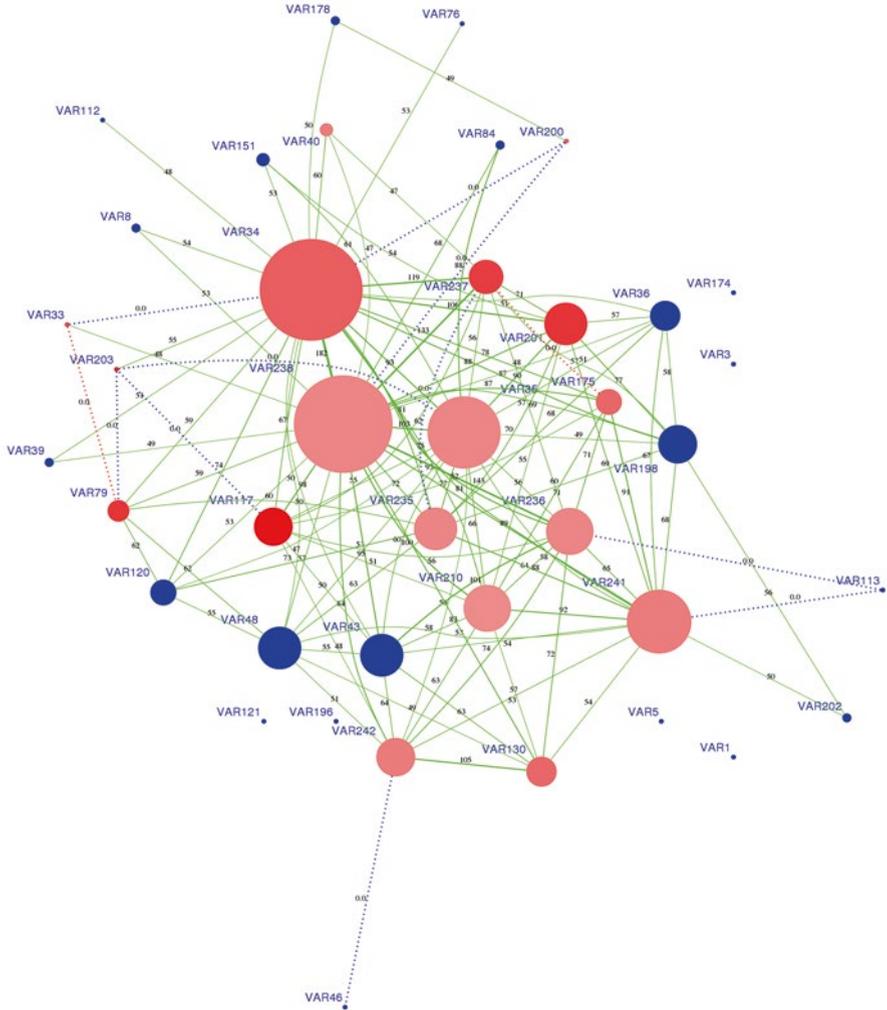


Fig. 7 Insomnia frequency network with superimposed prime implicants; *green edges* represent base frequency network; *red edges* represent prime implicant edges whereby the node pair results in a sub-additive effect; *blue edges* represent prime implicant edges whereby the node pair results in a super-additive effect. The node size represents degree of interaction in the base graph. *Blue nodes* represent nodes in the base network whereas the *red nodes* are prime implicant nodes. The degree of redness indicates frequency of presence in prime implicants (Figure 7 in high resolution with colour can be accessed at <http://www.sydney.edu.au/it/~itcm/book/images/figure2-7.jpg>)

correlation graphs are shown in Figs. 7 and 8. An energy layout is similar to the Kamada-Kawai algorithm was used for both core herb networks in Pajek (Batagelj and Mrvar 2003). This layout positions the graph such that the edge weights represent the edge length between two nodes. Higher frequency and correlated edges are therefore positioned closer together, and outlier frequency and correlation are further away from the primary network.

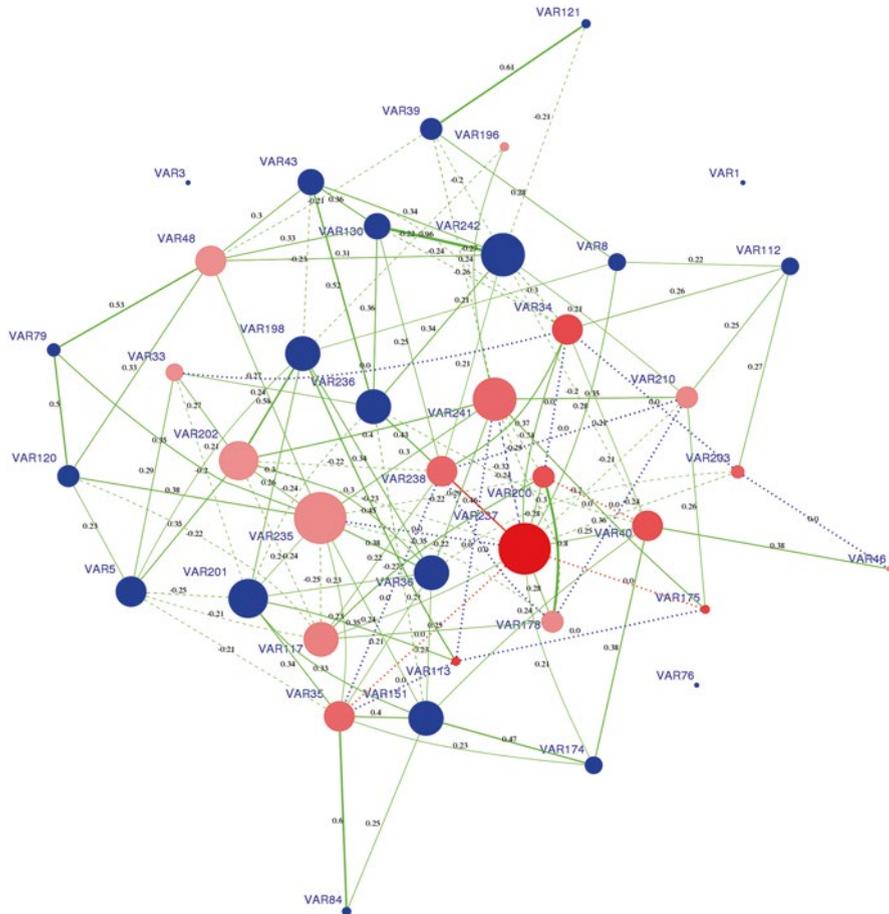


Fig. 8 Insomnia correlation network with super-imposed prime implicants; *green edges* represent base frequency network; *red edges* represent prime implicant edges whereby the node pair results in a sub-additive effect; *blue edges* represent prime implicant edges whereby the node pair results in a super-additive effect. Edges with a dashed style represent correlation < -0.2 , edges with a solid style represent correlation > 0.2 and edges with a dotted style represent correlation between -0.2 and 0.2 . The node size represents degree of interaction in the base graph. *Blue nodes* represent nodes in the base network whereas the *red nodes* are prime implicant nodes. The degree of redness indicates frequency of presence in prime implicants (Figure 8 in high resolution with colour can be accessed at <http://www.sydney.edu.au/it/~itcm/book/images/figure2-8.jpg>) (The full color version of this image may be viewed in the eBook edition)

6 Discussion

The previous section describes the methodology and demonstrates its application on the insomnia dataset. Thus, general observations can be made from visualization of the results, though further analysis is necessary to conclude its validity.

Although two base visualisation and analysis methodologies, frequency and correlation, have been introduced the complexity of herb-herb interactions may affect the validity of these inferences. In the illustrated network construction, negative NOT(herb) results were ignored. These confounding factors remain to be addressed in future work. Confounding in statistics is where a factor that may exert an effect is not measured in the experiment; thus biasing any analysis performed. Herbal combinations that appear in both positive and negative frequency and correlation networks indicate there may be possible confounding herbal or background factors that are affecting accurate analysis of core herbs. Confounding may arise through herb to herb interactions and extend to different and higher order herbal configurations, particularly when herbs are interacting amongst other herbs. Auxiliary herbs may also give rise to confounding due to their dependencies on other core herbs. While auxiliary herbs may not contribute to core therapeutic effects, their high frequency of usage may cumulate to significant effects. In complex disease patterns, the use of auxiliary herbs may exceed those of principal herbs. Further, though confounding herbs may be observed, this does not necessitate the removal of those herbs, which may further bias and confound the analysis. Addressing confounding and the absence of an herb can therefore be useful to illuminating more significant core herbs and less useful herbs.

7 Conclusion

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, especially for those herbs that are tightly connected to other herbs, as the later may impact negatively on the outcome. In this chapter, we have described and demonstrated an approach to discover the intertwining patterns of herbs in TCM prescriptions. Applying techniques in set-theoretic may distil the configuration where necessary herbs are required for the successful treatment. In addition, a network tool based on frequency of usage and correlation aided the understanding of domain knowledge choice of herbs as well as interesting factors that are worthwhile testing for. By further super-imposing positive outcome prime impli-cant results, a map of strong herb combinations with large positive outcome coverage can be inferred. This framework not only has validated results in consistence with earlier work performed by Zhou et al. (2011b), and also introduced a more efficient approach to reach configuration solutions form the causal complexity perspective.

Although this chapter introduces a computational approach for finding the useful herb combinations in the context of clinical outcomes. Several key issues still exist to be addressed in the future work. Firstly, two important information components, namely herb dosage and clinical manifestation (e.g. symptoms, co-morbid conditions), of the clinical data should be considered. Because it is widely recognized in the medical field that the herb dosage has an important effect for treatment

and also performs a significant role for complex interactions amongst herbs. Secondly, by incorporating dosage information, the computing cost of moving crisp-set QCA to fuzzy-set QCA to find the optimal complex herbal formulations should be further studied.

References

- V. Batagelj, A. Mrvar, *Pajek – Analysis and Visualization of Large Networks*. Graph Drawing Software, 3ed edn. (Springer, Berlin, 2003)
- D. Ehring, Causal asymmetry. *J. Philos.* **79**(12), 761–774 (1982)
- P. Fiser, J. Hlavicka, BOOM – a heuristic Boolean minimizer. *Comput. Inform.* **22**, 1001–1033 (2003)
- P.C. Fiss, A set-theoretic approach to organisational configurations. *Acad. Manage. Rev.* **32**(4), 1180–1198 (2007). doi:[10.5465/amr.2007.26586092](https://doi.org/10.5465/amr.2007.26586092)
- S.J. Hong, R.G. Cain, D.L. Ostapko, MINI: a heuristic approach for logic minimization. *IBM. J. Res. Dev.* **18**(5), 443–458 (1974). doi:[10.1147/rd.185.0443](https://doi.org/10.1147/rd.185.0443)
- T.K. Jain, D.S. Kushwaha, A.K. Misra, Optimization of the Quine-McCluskey method for the minimization of the Boolean expressions, in *Autonomic and Autonomous Systems, 2008. ICAS 2008. Fourth International Conference on*, 16–21 Mar 2008, pp. 165–168. doi:[10.1109/icas.2008.11](https://doi.org/10.1109/icas.2008.11)
- A. Kagliwal, S. Balachandran, Set-cover heuristics for two-level logic minimization, in *VLSI Design (VLSID), 2012 25th International Conference on*, 7–11 Jan 2012, pp. 197–202. doi:[10.1109/vlsid.2012.70](https://doi.org/10.1109/vlsid.2012.70)
- D. Katz, R.L. Kahn, *The Social Psychology of Organizations*, 2nd edn. (Wiley, New York, 1978)
- E.J. McCluskey, Minimization of Boolean functions. *Bell. Syst. Tech. J.* **35**(5), 1417–1444 (1956)
- S.K. Poon, K. Fan, J. Poon, C. Loy, K. Chan, X. Zhou, R. Zhang, Y. Wang, J. Xie, B. Liu, P. Kwan, J. Gao, D. Sze, Analysis of herbal formulation in TCM: infertility as a case study, in *Proceedings of the International Workshop on Information Technology for Chinese Medicine, In Conjunction with the IEEE International Conference on Bioinformatics & Biomedicine (BIBM 2011)*, Atlanta, 12–15 Nov 2011a.
- S.K. Poon, J. Poon, M. McGrane, X. Zhou, P. Kwan, R. Zhang, B. Liu, J. Gao, C. Loy, K. Chan, D. Sze, A novel approach in discovering significant interactions from TCM patient prescription data. *Int. J. Data. Mining. Bioinform.* **5**(4), 353–367 (2011b)
- W.V.O. Quine, The problem of simplifying truth functions. *Am. Math. Mon.* **59**(8), 521–531 (1952)
- W.V.O. Quine, A way to simplify truth functions. *Am. Math. Mon.* **62**, 627–631 (1955)
- C.C. Ragin, *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies* (University of California Press, Berkeley, 1987)
- C.C. Ragin, *Fuzzy Set Social Science* (University of Chicago Press, Chicago, 2000)
- C.C. Ragin, *Redesigning Social Inquiry: Fuzzy Sets and Beyond* (University of Chicago Press, Chicago, 2008)
- A. Su, S.K. Poon, J. Poon, Discovering causal patterns in TCM clinical prescription data using set-theoretic approach, in *Proceedings of the International Workshop on Information Technology for Chinese Medicine, In Conjunction with the IEEE International Conference on Bioinformatics & Biomedicine (BIBM 2013)*, Shanghai, 18–21 Dec 2013
- X. Zhou, S. Chen, B. Liu, R. Zhang, Y. Wang, P. Li, H. Zhang, Z. Gao, X. Yan, Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. *Artif. Intell. Med.* **48**(2-3), 139–152 (2010a)
- X. Zhou, J. Poon, P. Kwan, R. Zhang, Y. Wang, S.K. Poon, B. Liu, D. Sze, Novel two-stage analytic approach in extraction of strong herb-herb interactions in TCM clinical treatment of insomnia, in *Medical Biometrics*, ed. by D. Zhang, M. Sonka, vol. 6165 (Springer, Berlin/Heidelberg, 2010b)



<http://www.springer.com/978-3-319-03800-1>

Data Analytics for Traditional Chinese Medicine
Research

Poon, J.; K. Poon, S. (Eds.)

2014, XII, 248 p. 59 illus., 45 illus. in color., Hardcover

ISBN: 978-3-319-03800-1