HERBAL MEDICINES:

PHYSICIANS' RECOMMENDATION AND CLINICAL EVALUATION OF ST. JOHN'S WORT FOR DEPRESSION

ZORIAH AZIZ

Doctor of Philosophy

THE UNIVERSITY OF ASTON IN BIRMINGHAM

September 2003

HERBAL MEDICINES:

PHYSICIANS' RECOMMENDATION AND

CLINICAL EVALUATION OF ST. JOHN'S WORT

FOR DEPRESSION

The University of Aston in Birmingham

Herbal Medicines: Physicians' recommendation and clinical evaluation of St. John's wort for depression

Zoriah Aziz

Doctor of Philosophy 2003

Why some physicians recommend herbal medicines while others do not is not well understood. We undertook a survey designed to identify factors, which predict recommendation of herbal medicines by physicians in Malaysia. About a third (206 out of 626) of the physicians working at the University of Malaya Medical Centre were interviewed face-to-face, using a structured questionnaire. Physicians were asked about their personal use of, recommendation of, perceived interest in and, usefulness and safety of herbal medicines. Using logistic regression modelling we identified personal use, general interest, interest in receiving training, race and higher level of medical training as significant predictors of recommendation.

St. John's wort is one of the most widely used herbal remedies. It is also probably the most widely evaluated herbal remedy with no fewer than 57 randomised controlled trials. Evidence from the depression trials suggests that St. John's wort is more effective than placebo while its comparative efficacy to conventional antidepressants is not well established. We updated previous meta-analyses of St. John's wort, described the characteristics of the included trials, applied methods of data imputation and transformation for incomplete trial data and examined sources of heterogeneity in the design and results of those trials. Thirty randomised controlled trials, which were heterogeneous in design, were identified. Our meta-analysis showed that St. John's wort was significantly more effective than placebo [Pooled RR 1.90 (1.54-2.35)] and [Pooled WMD 4.09 (2.33 to 5.84)]. However, the remedy was similar to conventional antidepressant in its efficacy [Pooled RR 1.01 (0.93 –1.10)] and [Pooled WMD 0.18 (- 0.66 to 1.02).

Subgroup analyses of the placebo-controlled trials suggested that use of different diagnostic classifications at the inclusion stage led to different estimates of effect. Similarly a significant difference in the estimates of efficacy was observed when trials were categorised according to length of follow-up. Confounding between the variables, diagnostic classification and length of trial was shown by loglinear analysis.

Despite extensive study, there is still no consensus on how effective St. John's wort is in depression. However, most experts would agree that it has some effect. Our meta-analysis highlights the problems associated with the clinical evaluation of herbal medicines when the active ingredients are poorly defined or unknown. The problem is compounded when the target disease (e.g. depression) is also difficult to define and different instruments are available to diagnose and evaluate it.

Key words: predictors, logistic regression, meta-analysis, subgroup analysis, and heterogeneity

Acknowledgements

I thank Professor Alain Li Wan Po, my supervisor, for his constant interest, helpful guidance and kind assistance throughout the course of this research.

I am grateful to Professor Farouk El-Sabban for his helpful comments and contributions during the planning stage of the survey.

The scholarship from the Government of Malaysia is also gratefully acknowledged.

Thanks are also due to my friends Mick, Dr. Zhang, Dr. Khattab, Lubna, Ijeoma, Bernard, Paul and Ying for their friendships.

My heartfelt thanks must go to my family, Nadzri, Man and Min for their patience and understanding.

List of Contents

Appendices

List of Tables

List of Figures

Abbreviations

5-HT	Serotonin
CAM	Complementary and Alternative Medicine
CI	confidence interval
DSM	Diagnostic and Statistical Manual of Mental Disorders [versions include 1st edition (DSM-I), 2nd edition (DSM-II), 3rd edition (DSM-III), revised 3rd edition (DSM-IIIR), and 4th edition (DSM-IV)]
HAMD	Hamilton rating scale for depression
ICD	World Health Organisation International Classification of Disease [version include 6th edition (ICD-6), 8th edition (ICD-8), 9th edition (ICD-9), and 10th edition (ICD-10)]
MAOIs	monoamine oxidase inhibitors
NNT	number needed to treat
OR	odds ratio
RCTs	randomised controlled trials
RD	risk difference
RR	relative risk
SD	standard deviation
SE	standard error
SEM	standard error of the mean
SPSS	Statistical Package for the Social Sciences
SSRIs	selective serotonin reuptake inhibitors

TCAs	tricyclic antidepressants
UMMC	Universiti Malaya Medical Centre
USP	United States Pharmacopoeia
WHO	World Health Organisation
WMD	weighed mean difference
z	standard normal statistic
χ^2	Chi-squared statistic

CHAPTER 1

General Introduction

1.1 Complementary and Alternative Medicine

Complementary and Alternative Medicine (CAM) has increasingly become a focus of public attention and discussion (O'Connor et al., 1997). Yet there is no single universally accepted definition of CAM. Many writers have used Eisenberg et al.'s (1993) definition: "CAM are those treatments and health care practices not taught widely in medical schools, not generally used in hospitals, and not usually reimbursed by medical insurance companies". However, this definition is almost obsolete because courses on CAM are now being offered in many universities in the USA (Wetzel et al., 1998) and UK. In UK, for example, postgraduate studies in CAM are available at the University of Exeter.

More recently, the Cochrane Colloboration used the following more encompassing definition of CAM:
"A broad domain of healing resources that encompasses all health systems, modalities, and practices and their accompanying theories and beliefs, other than those intrinsic to the politically dominant system of a particular society or culture in a given historical period. CAM includes all such practices and ideas self-defined by their users as preventing or treating illness or promoting health and well-being. Boundaries within CAM and between the CAM domain and that of a dominant

system are not always sharp or fixed."

With this definition, herbal medicines also called phytomedicines or botanical medicines fall within the field of CAM. In addition to herbal medicines, CAM also includes a broad range of therapies and practices such as acupuncture, bio-electromagnetic therapy, reflexology, manual healing, homeopathy, mind-body interventions, nutritional aids and life-style changes.

1.2 Herbal medicines

Herbal medicines are the second most popular type of CAM after relaxation techniques (Eisenberg et al., 1998). They are different from the other CAM because their effectiveness can be evaluated using standard pharmacological approaches, the same way as standard pharmaceuticals (Levin et al., 1997). However, herbal medicines often have many components, which makes characterisation more complex than conventional single-agent pharmaceuticals. However, there is no difference in the research methods that can be used to test the effectiveness of herbal medicines or pharmaceuticals. Both are best evaluated by randomised controlled trials (RCTs). On the other hand, many of the other CAM are based on theories or concepts, which do not conform to current medical thinking (Eskinazi, 1998). Acupunctute, for example relies on a system of energy meridians. And homeopathy is based on the concept that the lower the dilution of a therapeutic agent (down to infinitely low dose), the more potent it is.

Herbal medicines as defined by The World Health Organization (WHO, 1996) can be classified into three categories:

a. Phytopharmaceuticals often sold as over the counter products in dosage forms such as tablets, capsules and liquids for oral use
b. Dietary supplements containing herbal products, also called nutraceuticals available in pharmaceutical dosage forms.
c. Phytomedicines, which consist of crude, semi-processed or processed medicinal plants.

The first and second category above are usually commercially packed and sold over the counter. Their primary active constituents are materials extracted or derived from natural plant sources. These two categories are popular and represent an area of great growth in the herbal industry because consumers in developed countries and those in urban areas of developing countries commonly use them. Consumers in the rural areas who rely on herbal medicines for their healthcare needs normally use products in the third category. These crude or semi-processed herbal medicines have an important place in primary health care in developing countries.

1.3 Evidence-based herbal medicines

The major problem with many herbal medicines currently available in the market is the lack of scientific evidence regarding their effectiveness and safety. Literature reviews, particularly those related to the historical use of such medicines, are of limited value and a significant number of such publications are oftentimes in languages that are less accessible than English. Additionally, many reputable medical journals do not publish studies on herbal medicines. This has been attributed to the lack of scientific evidence of their therapeutic actions and the poor methodology adopted in the studies (Buckman and Lewith, 1994).

Previous reviews on herbal medicines were of the "traditional" type. Such articles were prone to bias because reviewers selected the evidence, which they preferred. However, this type of review has become outdated (Ioannidis and Lau, 1998; Ioannidis et al., 1998). A systematic review aims to provide information about the effectiveness of interventions by defining precisely the review objective, retrieving all available evidence (defining inclusion and exclusion criteria), appraising the outcomes, and pooling any suitable quantitative data (meta-analysis) (Li Wan Po, 1998). Thus, a systematic review aims to be objective and to minimise bias.

Most of the currently available systematic reviews are on herbal medicines such as St. John's wort and Gingko biloba, which are widely used in developed countries. Many herbal medicines that are commonly used in non-developed countries have not been subjected to clinical research. Limited research is done on herbal medicines mainly because of the practical methodological problems when dealing with multiple herbal ingredients. In addition, many herbal medicines manufacturers are unable to invest in trials because of limited resources and expertise.

There is a strong push to apply the principles of evidence-based medicine to herbal medicines. For the medical community to feel comfortable in recommending a particular herbal product, acceptable evidence of its safety and efficacy must be generated. Unfortunately, there are few high quality clinical trials of those products available. Much of the research on herbal medicines has been conducted in Asia and continental Europe. As such the results are often not reported in English. Because of these two factors namely the scarcity of quality trials and the difficulty in accessing the existing data, it is important to summarise what is available systematically.

In evidence-based medicine, recommendations of treatments are based on the quality of evidence available. Evidence is organised in levels, which reflect the quality of the design and methodology of the trials. In order to assess the evidence when considering botanical medicines for inclusion in the United States Pharmacopoeia (USP), the USP has adopted an assessment based on four levels of evidence, (USP Press Release, 2000), as summarized below. Level I include trials of the highest quality and level IV, anecdotal evidence.

USP criteria for Levels of Evidence

Level I	Randomised controlled trials, meta-analyses and epidemiological studies of highest quality
Level II	Randomised controlled trials, meta-analyses and epidemiological studies of moderate quality
Level III	Inconclusive studies
Level IV	Anecdotal Evidence

Adapted from (USP Press release, 2000).

As with any grading systems, there are limitations particularly with respect to the quality of the trials. Nonetheless, using the USP system can provide an initial approach to objectively evaluate the quality of evidence supporting use of specific herbal medicines. For example, based on this system, there are some studies, which can be graded as having Level II evidence. Examples of such studies are, randomised controlled trials evaluating the effectiveness of Chinese herbal medicines for irritable bowel syndrome (Bensoussan et al., 1998), feverfew for prevention of migraine (Murphy et al., 1988), gingko biloba for dementia (Le Bars et al., 1997).

However, with the exception of St. John's wort for depression (Linde and Mulrow, 2003), most herbal medicines are supported by weaker levels of evidence. The evidence of their effectiveness is only based on testimonials or hearsay (Barret et al., 1999), which is Level IV evidence. Generally the trials of these medicines have flawed study designs such as small sample size, undefined outcome measures, failure to account for biases, poorly defined exclusion or inclusion criteria and use of non-standardised products.

1.4 Specific problems with herbal medicines

In contrast to other herbal medicines, the efficacy of St. John's wort for depression has been widely investigated in controlled trials (Linde et al., 1996) and it has reached level 1 of evidence according to USP criteria for levels of evidence (USP Press Release, 2000). Investigations on the effectiveness of St. John's wort have illustrated that evidence- based principle can be applied to herbal medicines. However promising this evidence may be, when dealing with herbal medicines, interpretation is difficult by the fact that herbal medicines have specific problems.

Unlike conventional medicines, which contain known chemical substances, they are complex remedies with unknown chemical entities. Also, in contrast to conventional drugs, which are labelled according to their International Non-proprietary Names, advocated by the WHO, the nomenclature for herbal medicines can be confusing. Apart from the botanical names of the herbs, local names, which are less specific/accurate and vary from region to region, are often used.

Additionally, most drug regulatory agencies do not apply stringent quality control measures to herbal medicines, as they do for pharmaceuticals. As a result, the ingredients listed on the label may not be present in the stipulated amounts. The same herbal medicines may have variable compositions (Cui et al., 1994). This is also illustrated with St. John's wort products (Busse, 2000). Although the hypericin content was between 0.18% and 0.25% in products standardized to this chemical, the supposed active constituent, hyperforin, varied from 0% to 3.26%.

Another problem related to lack of regulation is with regards to unsubstantiated claims of benefits made. The lack of published evidence does not deter unscrupulous manufacturers or distributors from making unsupported and exaggerated therapeutic claims to the unsuspecting public. In fact, consumers may be fooled into believing that such claims are true since there is no evidence to state otherwise. Unfortunately, serious conditions such as cancer, obesity, HIV infection and heart disease are often the targets of unsubstantiated claims, with desperate patients as easy preys.

Given that herbal medicines are often portrayed as being harmless, the public may be at risk of adverse effects from the use of herbal medicine. A lack of good quality information on herbal products compounds the problem. Literature reviews, particularly those related to the historical use of such medicines, are of limited value and a significant number of such publications are often in difficult to access foreign language journals. A lack of good scientific evidence on herbal medicines and studies of poor methodological quality means that few articles are published in the mainstream scientific journals (Buckman and Lewith, 1994).

Blinding is another specific issue in clinical trials of herbal products. Given the nature of herbal products, blinding participants and researchers in clinical trials of herbal medicines is not easy (Gaus and Hogel, 1995).

1.5 Aims of Study

The work focused on two aspects of herbal medicines.

The first part deals with a survey carried out in a teaching hospital in Malaysia. The aim was to identify variables that were predictive of physicians' recommending herbal medicines to patients.

The second part concerns a critical evaluation of the evidence-base for St. John's Wort for depression. The specific aims were to:

a. describe the characteristics of randomized controlled trials of St. John's wort for depression.
b. apply methods of data imputation, approximation and transformation when reported data from trials are incomplete
c. undertake a meta-analysis of trials of St. John's wort for depression
d. evaluate sources of heterogeneity in randomized controlled trials of St. John's wort.

PART I: SURVEY OF PHYSICIANS IN A TEACHING HOSPITAL IN MALAYSIA

CHAPTER 2

Herbal medicines: predictors of recommendation by physicians

2.1 Introduction and aim

Herbal medicine, also called phytomedicine or botanical medicine, is the second most popular type of Complementary and Alternative Medicine (CAM) after relaxation techniques (Eisenberg et al., 1998). Whilst, there are numerous surveys examining the use and practice of CAM in various general populations (Eisenberg et al., 1993; MacLennan et al., 1996; Nicassio et al., 1997; Astin, 1998; Bausell et al., 2001; Cherniack et al., 2001) and among physicians (Marshall et al., 1990; Knipschild et al., 1990; Borkan et al., 1994; Berman et al., 1995; Ernst et al., 1995), there have been relatively few surveys on the practice of herbal medicine.

In their review of 19 surveys, which examined the practices, and beliefs of conventional physicians with regard to 5 of the more prominent CAM therapies, Astin et al. (1998) reported that the extent to which physicians practised herbal medicine varied considerably across countries and cultures. For example, a Scottish study found that none of the physicians surveyed practised herbal medicine (Reilly, 1983). A corresponding figure of 2% was found in a study of physicians in Auckland, New Zealand (Marshall et al.,

1990) and one of 23% in a US study by Blumberg et al. (1995). The highest figure (78%) was seen in the study of physicians in Kassel, Germany (Himmel et al., 1993).

Factors which have been identified to be associated with physician recommending CAM or referring patients to CAM practitioners are gender (Verhoef and Sutherland, 1995; Burg et al., 1998; Berman et al., 2002; Corbin-Winslow and Shapiro, 2002), length of practice (Berman et al., 1998), practice type, location of medical training (Verhoef and Sutherland, 1995; Sikand and Laken, 1998), interest (Berman et al., 1995), and beliefs in the legitimacy of the CAM therapies (Berman et al., 2002). Additionally, Berman et al. (1998) found that knowledge of a therapy, which is measured through training, also predicts CAM practice.

Herbal medicine is a booming industry in many countries including Malaysia. Although currently there is no data in Malaysia on the prevalence of use among the general population, the herbal market is viewed as an emerging sector with big commercial opportunities by manufacturers. The annual sales of traditional medicines, the greater proportion of which comprises of herbal medicines, has been reported to be RM1 billion (US$ 260 millions) in Malaysia (MIGHT, 2002).

The increasing use of traditional medicines, including herbal medicines, raises significant public health issues, including safety. It was this need to protect the public that led to the regulation of traditional medicines through legally enforced product registration in Malaysia in January 1992. By December 2000, slightly more than 20,000 applications for registration had been submitted to the Drug Control Authority (DCA), Malaysia. Of this, only half have been approved for registration (Annual Report NPCB, 2000). However, the review of herbal medicines is not as rigorous as that of synthetic and more conventional medicines. The approval process is simplified and marketing authorisation is granted if the herbal medicine is not promoted and labelled for preventing or treating any of the twenty diseases in the List of Prohibited Claims, as specified in the Medicines (Advertisement and Sale) Act 1956 – (Revised 1983). Examples of those diseases include diabetes, cancer and mental disorders. To get round this legal provision, manufacturers sometimes use promotional materials to hint at the prohibited claims. This unethical practice is facilitated when labels carrying statements such as 'promotes prostate health' and 'maintains proper reproductive function' are permitted. Lack of human resources and the expanding scope of regulatory control, make it difficult for any regulatory authority, including the Ministry of Health Malaysia, to identify every violation of the advertisement or sales regulations.

Physicians tend to be more demanding than the general public regarding evidence for efficacy (Eisenberg et al., 1993). This greater demand for evidence together with the apparent increasing popularity of herbal medicine, the inadequate regulatory structure for controlling them and the paucity of objective evidence for most herbal medicines, lead to real problems for physicians, healthcare adviser and patient alike. We hypothesised that as the use of herbal medicine increases in the general population, so do patients' request to physicians for herbal medicine recommendations. Some physicians are likely to respond to patients' requests. However, little is known about which characteristics of physicians are predictive of increasing the likelihood of recommendation of herbal medicines to patients.

The aim of this study was to better understand herbal medicines in general but more specifically to identify factors, which predicted physician's likelihood of recommending herbal medicines. An understanding of these is important in order to understand how physicians advise patients and to identify medical training needs of physicians.

In this study the term herbal medicines is used to describe all herbal products, including herbal extracts, that are commercially packed and sold over the counter.

2.2 Methods

2.2.1 Design and sample

The design of the study was a convenient sample of physicians working in University Malaya Medical Centre (UMMC). This is a teaching hospital in the capital city of Malaysia, Kuala Lumpur. Physicians across all specialities were represented in the survey. Face-to-face interviews, using a structured questionnaire (Appendix 1) were conducted over a one-month period in February 2001.

The first draft of the questionnaire used in the survey was tested on 20 physicians from the UMMC. Data from the pilot study was used to adjust the questionnaire and were not included in the final analysis. The study protocol was approved by the Director of UMMC.

2.2.2 Survey instrument

The survey instrument consisted of five sections:

a. View on interest, usefulness and safety of herbal medicines

This section assessed physicians' interest and views on the usefulness and safety of herbal medicines. Responses were collected using a 5-point Likert scale.

b. Opinions on herbal medicines

This section sought the physician's opinion on herbal medicine. Since opinion is a broad construct, four measures were used to tap it:

1) Herbal medicines should only be used in the treatment of minor health problems (e.g., common colds and coughs)
2) Herbal medicines should NOT be used for serious health problems (e.g., chronic asthma, stroke).
3) Herbal medicines should only be used if there is evidence of effectiveness from randomised clinical trials.
4) Scientific evidence required for herbal medicines must be similar to that of conventional medicines.

Using a five-point Likert scale ('strongly disagree'= 1, 'disagree'= 2, 'neither disagree or agree'=3, 'agree'= 4 and 'strongly agree'= 5), the physicians were required to rate their agreement or disagreement with each of the four statements which were used for the construction of the opinion scale.

c. Personal use and practise (recommendation to patients)

Data were collected on whether physicians personally used herbal medicine and whether they had recommended herbal medicines to their patients.

d. Training

This section assessed respondents' previous training in herbal medicines, interest in receiving training on the subject and whether they approved of incorporation of teaching of herbal medicine in the medical curriculum.

e. Demographic characteristics

Physicians' recommendation of herbal medicine may be influenced by personal characteristics. Thus demographic data collected were age, gender, ethnicity and level of medical qualification (both basic and higher) and number of years in practice.

2.2.3 Data analysis

Data were coded and entered into spreadsheets (Microsoft Office Excel ®) and subjected to statistical analysis using Statistical Package for the Social Sciences (SPSS), version 10. Logistic regression modelling was used to assess the significance of potential predictor variables on recommendation of herbal medicines by the physicians.

Model building

Table 2.1 lists the independent variables considered as possible predictors of recommendation of herbal medicines by physicians.

Table 2.1 Possible predictor variables

a. Personal Use

b. Interest and perception of usefulness and safety of herbal medicines

c. Opinion related to use and views on type of evidence

d. Training, including interest in receiving training and agreement to incorporation of teaching of herbal medicine in the medical curriculum.

e. Demographic information; gender, race, age, level of medical training and years in practice.

Personal use, interest in receiving training and agreement to incorporation of teaching of herbal medicine in the medical curriculum were regarded as dichotomous variables and coded 0 = "No" response and 1 = "Yes" response. Similarly, gender (male; female), age (<36 years; 36 years and older) and level of medical training (basic; higher) were coded arbitrarily as 0 = for the first level and 1= for the second level. Nonordered categorical data with more than two levels such as race was entered as k-1 dummy variables with Malay race as the reference group and k equal to 4 in the present study.

2.3 Results

Two hundred and fifteen physicians from a total of the 626 physicians working at the Centre were approached for the face-to-face interview using the structured questionnaire. Of this, 206 (95.8%) agreed and consisted of 57 per cent males and 43 per cent females. Ethnic Chinese physicians made up 39 per cent of the respondents, followed by Malay (33%), Indian (21%) and the others 7 per cent. The majority of the respondents (75.2%) had post-graduate medical training.

Using principal components analysis with varimax rotation on the four statements, 2 item factors were identified from the responses probing physician's opinions. We named OPINION 1 (items related to the use of) and OPINION 2 (items related to evidence). The interpretation of these two factors was consistent with two statements loading on appropriateness of use and the quality of evidence.

Redundancy was examined with correlation analysis of the predictor variables. No serious multicollinearity was observed. However, a moderate correlation was obtained between the variable INTEREST and PERSONAL USE [r=0.49; p = 0.01 (2-tailed)]. To obtain the most parsimonious model, backward stepwise regression analysis was undertaken, starting with the full main effects model and including only those that are statistically significant at the 0.05 significance level. None of the variables excluded on this basis were judged to be essential to force into the model.

The final model showed that respondents, who personally used herbal medicines, were more likely than non-users to recommend herbal medicines, Odds ratio of 3.8 (95%CI, 1.5-10.2) (Table 2.2).

The "odds ratio" for the RACE (Chinese), RACE (Indian) and RACE (Others) coefficients were less than 1. This implies that Malay physicians were more likely to recommend herbal medicine compared to physicians from other races such as Chinese, Indian and others (non-Malaysian nationals). The odds of Malay physicians recommending herbal medicine were four times that of the Chinese physicians (OR=0.24), and about five times that of the Indian physicians (OR=0.20).

Table 2.2 Predictors of herbal medicines recommendation in the multivariate logistic regression model

Variables	**Beta**	**SE (beta)**	**Odds Ratio**	**95 % CI**
PERSONAL USE	1.346	0.498	3.84	1.45 to 10.19
INTEREST	0.968	0.312	2.63	1.43 to 4.85
INTEREST (TRAINING)	2.695	0.836	14.80	2.87 to 76.21
RACE (Chinese)	-1.409	0.516	0.24	0.09 to 0.67
RACE (Indian)	-1.635	0.655	0.20	0.05 to 0.70
RACE (Others)	-3.000	1.221	0.05	0.01 to 0.55
HIGHER TRAINING	1.950	0.638	7.03	2.01 to 24.55

Variable for RACE used Malay as the Reference group.

Physicians with an INTEREST in training had odds of 14.80 with a 95% confidence interval of (2.87, 76.21), relative to that of physicians who did not, other factors being constant. Similarly, those who had higher level of medical training had odds of 7 times that of physicians without such training, for recommending herbal medicine.

2.4 Discussions

The study found that about one in five (19 per cent) of the respondents recommended herbal medicines. Significant predictors were interest in herbal medicine, interest in receiving training in this area, personal use of herbal medicines, higher level of medical education and being a Malay physician. We found that Malay physicians were more likely to recommend herbal medicines than Chinese, Indian or physicians of other nationalities.

The prevalence of physicians recommending herbal medicines (19%) found in this study is considerably higher than the 3.6% reported in a US study (Jeffrey et al., 1998) but much lower than the 78% reported in a German study (Himmel et al., 1993). One possible explanation for these variations is cultural differences in the acceptance of herbal medicine. Danesi (1993) demonstrated, that culture, which includes ethnicity, practices, beliefs and values, influenced cultural perceptions of 'health' and 'disease'. As such, the extents to which physicians recommend herbal medicine may be influenced by culture. However time differences and differences in the type of physicians sampled may also be contributory.

Studies assessing attitudes towards herbal medicine used different measures of attitudes. For example, White et al. (1997) assessed attitudes by asking respondents to rate the effectiveness of the individual CAM therapies on a visual analogue scale, while Berman et al. (1995) asked respondents if they considered specific therapies as a legitimate medical practice. In contrast, Sikand and Laken (1998) used three variables associated with the CAM used to measure attitude. Since, the traditional method of measuring attitudes is by means of attitude statements (Oppenheim, 1996), we attempted to measure attitudes by using four statements associated with herbal medicine. However, factors that we named as OPINION 1 and OPINION 2 were not found to be significant predictors of herbal medicine recommendation among physicians. As attitudes are made up of several components of values, beliefs and feelings (Oppenheim, 1996), we had probably measured only part of attitude. Thus, future research should focus on developing a broader scale.

Physician's level of medical education predicted recommendation of herbal medicines. Surprisingly those with higher-level training in conventional medicine were more likely to recommend herbal medicines. One possible interpretation for this unexpected finding is that physicians with the higher level of qualification may be more tolerant of uncertainty and were more open to try out new approaches such as herbal medicine.

Our results are also consistent with those of other studies (Corbin-Winslow and Shapiro, 2002; Rooney et al., 2001), which found that physicians who personally used herbal medicines were more likely to recommend them than physicians who were not.

The results of this study indicate that, general interest as well as interest in receiving training in herbal medicine predicted physicians' likelihood of recommending herbal medicines to patients. These findings are in agreement with those of Berman et al. (1998) and Verhoef and Sutherland (1995). We did not explore the reasons for physicians' interest in receiving training, but we can only speculate that they wanted the training so that they could make appropriate recommendations. Contrary to previous studies (Burg et al., 1998; Berman et al., 1998; Berman et al., 2002; Corbin-Winslow and Shapiro, 2002), being female and the length of time the physicians were in practice, were not significant predictors for recommendation of herbal medicines.

The survey was conducted in a teaching hospital in an urban area, and therefore the results may not be generalisable to physicians working in other settings in Malaysia. Thus, extrapolation of our data to other non hospital-based physicians or other non-teaching hospitals in other regions of Malaysia or to other national settings should be made with caution. Perkin et al. (1994) found that the practice of and belief in CAM to be greater among general practitioners than among hospital-based physicians. Hence, future studies could seek to explore any differences with respect to both non-hospital based physicians and physicians working in other parts of the country.

2.5 Conclusion

Within a Malaysian urban teaching hospital setting, personal use, general interest, interest in receiving training, race and higher level of conventional medical training were significant positive predictors of physician recommendation of herbal medicines to patients. While the study identified factors, which have been shown to be associated with recommendations of herbal medicines, further research is needed to validate and replicate our findings. The development and validation of a broader instrument to further probe the attitudinal characteristics of physicians with respect to herbal medicine may also be worthwhile.

PART II

Clinical evaluation of St. John's wort for depression

CHAPTER 3

Depression and St. John's wort

3.1 Depression and prevalence

The clinical diagnosis of depression is made on the basis of the existence of a collection of signs and symptoms (syndrome) characterised by mood disturbance ranging from mild to severe and from transient to persistent (Preveler et al., 2002). Depression is one of the commonest conditions encountered in psychiatry and it is an illness, which is very disabling.

Depression severity is often simply defined as 'mild', 'moderate' or 'severe'. This term describes the extent to which the patient's everyday life is affected. For example mild depression causes only minor impairment of the patient's work, social life and relationships with others. As such major depression can be of mild severity. Major depression is a diagnostic category in the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (American Psychiatric Association, 1994). Moderate depression is associated with more obvious symptoms and is likely to be noticeable to others. Severe depression produces symptoms that affect the patient so badly that he or she may be unable to work or to relate socially to others.

Major depression is common, with a prevalence of between 5 and 10 percent of people seen in primary-care settings (Katon and Schulberg, 1992). Two to three times as many people may have depressive symptoms but do not meet criteria for major depression. Studies of major depression have consistently demonstrated that this disorder occurs 1.5 to 3 times more frequently in women than in men in the adult population. This disparity is consistent across cultures, samples, and assessment techniques (Weissman and Klerman, 1977; Weissman et al., 1996; Kessler et al., 1994). The WHO estimates that depression will soon be the second leading cause of disability worldwide after heart disease (Murray and Lopez, 1996). The diagnosis of depression is a challenge because it presents in various forms, often resembling other physical conditions as well as other mental illnesses.

The three types of therapy for depression which have proven efficacy are pharmacotherapy (drug treatments), psychotherapy and electroconvulsive therapy (William et al., 2000; Thase et al., 1997; Persad, 1990)

3.2 Drug treatments for depression

Researchers believe that depression is caused by one or more biochemical imbalances. It is thought that one such imbalance is of the neurotransmitters serotonin, dopamine, and norepinephrine. They are secreted into the synapses, or spaces, between neurons and are "taken up" by receptors where they are subsequently stored or metabolised with the aid of monoamine oxidase.

Compounds that interfere with or inhibit this process have been found to have a beneficial effect on depression.

Drug treatments for depression, falls into three major classes: tricyclic antidepressants (TCAs), monoamine oxidase Inhibitors (MAOIs), and selective serotonin reuptake inhibitors (SSRIs).

Some of the TCAs now available are amitriptyline, imipramine, maprotiline, desipramine and doxepin. They are still widely used as the first-line antidepressants. TCAs enhance the concentration of serotonin and norepinephrine by blocking their reuptake so that more of these neurotransmitters are available for the transmission of electrical impulses.

The MAOIs, which include isocarboxazid and phenelzine, are not as widely used as TCAs. They act by inhibiting monoamine oxidase, thereby slowing down neurotransmitter (serotonin and catecholamines) degradation.

SSRIs are the newer class of antidepressants, and include fluoxetine, paroxetine and sertraline. Their mechanism of action is the selective inhibition of serotonin (5-HT) reuptake from synapses.

Bupropion, which is chemically unrelated to the other antidepressants, inhibits the uptake of norepinephrine and dopamine, more than serotonin, and does not inhibit monoamine oxidase (Physicians' Desk Reference).[37]

3.3 Herbal remedies for depression

St. John's wort (*Hypericum perforatum*), valeriana (*Valeriana officalis L*), and kava kava (*Piper methysticum*) have all been used for depression. St. John's wort accounts for about 10% of all herbal medicine sold in the United States and has been called by some as "nature's own Prozac" (Rey and Walter, 1998). St. John's wort appears more effective than placebo for the short-term treatment of mild to moderately severe depressive disorders (Linde et al., 1996). Adverse effects are reported to occur significantly less frequently with St. John's wort compared to first generation tricyclic antidepressants (Linde et al., 1996).

3.4 Diagnostic classification of depression

The classification of depression is a controversial topic that has caused much debate in the field of mental health. Dispute arises because diagnoses are not based on specific physiological or laboratory test. Instead, the diagnoses are based on the signs and symptoms observed and identified. As such, making a diagnosis of depression is not always straightforward. Another problem is that different psychiatrists often use different criteria to diagnose depression. Much of the resulting confusion has however been curtailed by the availability of two internationally recognised sets of diagnostic criteria namely the Diagnostic and Statistical Manual (DSM) and the International Classification of Disease (ICD). These criteria have led to a greater uniformity in the diagnosis and classification of depressive illnesses.

Diagnoses and classification are useful to both research community and clinicians. Classifications of depressions are vital for communication, guiding treatment or prognosis and applying the results of research studies to clinical work.

In 1948, WHO published the sixth edition of ICD (ICD-6) that for the first time included a section on mental disorders. In 1952 a variant of ICD-6 was developed as the first edition of the Diagnostic and Statistical Manual: Mental Disorders (DSM-1). The next coordinated revision led to ICD-8 and DSM-II. Subsequently, another coordinated revision led to ICD-9 published in 1978 and DSM-III published in 1980. Because of a number of inconsistencies and unclear criteria in the DSM-III, it was then revised to DSM-IIIR which was published in 1987, followed by the latest version DSM-IV in 1994. The latest version for the ICD classification, which was published in 1993, is the 10th revision of the International Classification of Diseases (ICD-10).

3.4.1 DSM classifications

Major depressive disorder as defined by the DSM-IIIR is the broadest definition of depression (Appendix 2). The classification requires 5 (out of 9) depressive symptoms occurring together during at least a 2-week period, whereas major depressive disorder as defined by the DSM-IV further requires a clustering of 5 depressive symptoms during a period of 2 weeks or longer and additionally requires either depressed mood or a diminished interest or pleasure in daily activity, which occurs most of the day in a 2-week period (Appendix 3).

Depression may range in severity from mild symptoms to more severe forms that include delusional thinking, excessive somatic concern, and suicidal ideation, over longer periods of time. The depression can be further specified by the number of episodes (single versus recurrent).

The term "minor depression" has also been used to describe depressive conditions that are not of sufficient severity and duration to meet criteria for a major depressive episode. Additionally, DSM-IV classification of minor depression needs the absence of any previous depressive episode (Depression Guideline Panel, 1993).

3.4.2 ICD Classifications

The classification of ICD-9 is based on a system that differentiates between depression of endogenous, psychogenic or somatogenic origin and depressive personality disorders. The new ICD-10 offers the advantage of diagnoses based on operationally defined parameters of the depressive conditions, such as its signs and symptoms, its severity and its course. Diagnosis of depressive disorder by the ICD-10 system of classification requires assessment of the three listed characteristics of the disorder:

a. The signs of the condition (Appendix 4)
b. The severity (intensity) of the condition (Appendix 4)
c. The course (single episode, recurrent, duration) of the condition

Depending on the number and severity of the symptoms, a depressive disorder may be specified as mild, moderate or severe (Appendix 4). The ICD-10 system is most widely used throughout Europe and the UK.

3.4.3 Differences between ICD-10 and DSM-IV definitions

The ICD-10 and DSM-IV criteria for major depression consist of a similar collection of symptoms. These systems only describe the symptomatology of the illness and do not score the severity or intensity of illness. However, there are several important differences between the ICD-10 and DSM-IV definitions.

First, ICD-10 has a greater number of core symptoms and specifies that two out of three of these core symptoms (depressed mood, lost of interest or pleasure, fatigue) must be present for a diagnosis, whereas in DSM-IV only one out of two (depressed mood, lost of interest or pleasure) is required. This difference in the number of core symptoms results from the inclusion of fatigue among the core symptoms in ICD-10 (Appendix 4), whereas it is among the associated symptoms in DSM-IV (Appendix 3).

Second, ICD-10 requires at least one additional associated symptom for a total of four symptoms to diagnose depression, whereas DSM-IV requires a total of five symptoms.

Other differences in the two definitions of depression include the DSM-IV criteria requiring the presence of "clinically significant" distress or impairment and specifying an exclusion for bereavement, as well as the ICD-10 inclusion of an associated symptom of "loss of confidence or self-esteem."

However, Gavin et al. (1999) has demonstrated that these differences in the two classification systems do not produce a high number of discrepant diagnoses, a reflection of the similarity in diagnostic criteria.

In DSM IV, a major depressive episode is described in terms of severity and whether or not psychotic features are present. A major depressive episode may be mild, moderate or severe, according to the level of incapacitation of the patient's function at work or at home. In its mild form, major depression has a score of 7-17 on the Hamilton Depression Rating Scale: a score of 18-24 for moderate severity and a score of above 25 for severe episode (Snow et al., 2000). However, there may be some overlap with the scores making diagnosis of severity difficult.

3.5 Rating scales for assessing efficacy of antidepressants

There is a wide range of methods of assessing the efficacy of antidepressants with a confusing range of acronym. However, the essential research tool used in depression is rating scales. There are now numerous rating scales. These scales differ according to the observer, number of items, criteria used in the scaling of scores and the symptoms investigated (Hughes et al., 1982). Although most commonly used rating scales have been validated, they often place emphasis on different symptoms and factors, thereby rendering accurate comparisons among study results difficult (Faravelli et al., 1986). The use of standardised scales is important in comparing results between studies or summing up results in meta-analysis. There are many ways of classifying scales, but the most important differentiates between observer-rating scales and self-rating scales (Hamilton, 1976). Both types of scale provide valuable information and clinical studies sometimes use both scales to assess efficacy. Generally, experts in psychiatry agree that observer rating should be employed, as the primary measure and self-rating scales should be used to provide additional information (Möller, 2000).

The most commonly used observer-rating scale is the Hamilton Rating Scale for Depression (HAMD) (Hamilton, 1960). In contrast with the HAMD, in self-rating scales, patients complete the questionnaire. The self-rating scales currently in use, include the Beck Depression Inventory (Beck et al., 1979), the Zung Self-Rating Depression Scale (Zung, 1965), and the Center for

Epidemiological Studies Depression Scale (Radloff, 1977) and the Carroll Rating Scale (Carroll et al., 1981). The Beck Depression Inventory has been frequently used as an outcome measures in studies of depression. However, all these self-rating scales are seldom used to assess antidepressive efficacy, they are more commonly used as screening tools in research and clinical setting.

3.5.1 Hamilton Depression Rating Scale (HAMD)

First introduced by Max Hamilton in 1960, HAMD has since become the most widely used and accepted standard for evaluating the efficacy of antidepressant treatments (Snaith, 1996). However, it cannot be used to establish a diagnosis (Hamilton, 1967; 1980). The scale is also the normal standard against which other depression rating scales are validated (Carroll et al., 1981; Montgomery and Asberg, 1979). The HAMD scale which contained 17 scored items (variables) is an observer rating scale and provided only general guidelines for the administration and scoring of the scale. A number of investigators (Endicott et al., 1981; Miller et al., 1985; William, 1988) have modified the scale and made recommendations about administration and scoring procedures. Since then several versions of the scale with different numbers of items have been in use. Hedlund and Vieweg (1979) commented on the difficulties in achieving inter-rater reliability across trials settings caused by the varied versions of the scale. This makes comparison of results from different studies difficult especially when the version used is not reported.

The HAMD scale (Hamilton, 1960) had in addition to the 17 scored items, four additional items, which were not scored. Some investigators have extended scoring to all 21 items. The first 17 items of the 21-item HAMD are commonly used, as the final four items relate to diurnal variation, derealisation, paranoia, and obsessions, and were suggested by Hamilton (1960) to be of lesser prevalence and importance in assessing depression (Hamilton, 1960). Furthermore, a 24-item version adds symptoms of hopelessness, helplessness and worthlessness. Each item is rated on a 0 to 2 (three-points) or a 0 to 4 (five-points- for items that are more difficult to quantify) scale, and the scores are summed. The higher the score, the more severe is the depression.

Clinical trials of St. John's wort often include only patients with HAMD scores of 15 to 18 on the 17-item scale. Such patients are considered to have mild to moderate depression. Patients with severe depression typically score 25 or higher on the HAMD (Hamilton, 1960; Snow et al., 2000).

The HAMD Rating Scale is not ideal for older people because it includes a number of somatic items that may be positive in older people without depression. However, it is still the most widely used scale, even in studies involving the elderly, although specific scales, such as the Geriatric Depression Scale which exclude somatic items are available (Hope, 2003).

3.6 Outcome measure in depression

Many trials use continuous scales to measure depressive symptoms such as the HAMD Rating Scale and the Beck Depression Inventory. Changes in those continuous measures are then usually dichotomised. Response, typically defined as more than a 50% decrease from baseline score for a standardised scale (e.g. HAMD) is one of the most consistently used criteria to define improvement (Depression Guideline Panels, 1993).

3.6.1 Dichotomous data

The response rate, is usually defined as the proportion of patient having either (a) 50% or greater improvement from baseline as assessed by the rating scale or (b) a total score below a predetermined level, usually a score of 10 or less on the HAMD rating scale. Results are expressed as relative risks or odd ratio. Determining response on the basis of proportionate decrease in baseline score implies that such a responder is free from depression, which is not necessarily the case. For example, a severely depressed patient with a baseline HAMD score of 32 may have a 50% reduction in score to 16, which is still indicative of depression.

Although it is easier to understand results presented as a dichotomous data, but information is lost when continuous data are transformed to dichotomous data.

3.6.2 Continuous data

Results are often reported as change-from-baseline (the mean difference between pre and post treatment HAMD measurements), final HAMD score or rate of change of HAMD.

Montgomery (1994) has suggested that a difference of just four points on the HAMD score between treatment and control group may be important. However, there is no experimental work supporting Montgomery's figure or any other figure, and there is still no agreement about what represents a clinically significant improvement.

3.7 St. John's Wort

3.7.1 Uses

St. John's wort *(Hypericum perforatum)* is an herb that has been used for centuries to treat a range of ailments such as excitability, neuralgia, anxiety, sleep disorders, and depression (Newall et al., 1996). Today, it is best known for its use in depression and it is one of the top-selling herbal medicines. Brevoort (1998) reported that the sales of St. John's wort had increased 280 fold in just one year. It is often referred as "Nature's Prozac." (Rey and Walter, 1998; Schempp et al., 1999). The composition of St. John's wort and how it might work are not well understood. There is some scientific evidence that St. John's wort is useful for short-term treatment of mild to moderate depression (Linde and Mulrow, 2003). However, recent studies suggest that St. John's wort is of no benefit in treating major depression of moderate severity (Shelton et al., 2001; Hypericum Depression Trial Study Group, 2002).

3.7.2 Constituents

At least 10 constituents of St. John's wort extracts have been identified and could have contributed to its antidepressant activity including hypericin, pseudohypericin, hyperforin, and adhyperforin (Jensen et al., 2001; Müller 2003). The precise mechanism of action of St. John's wort is not yet clearly established. St. John's wort extract is believed to increase the availability of monoamine neurotransmitters in the synaptic cleft by inhibiting the reuptake of serotonin, norepinephrine, and dopamine in a manner similar to that of conventional antidepressants (Muller and Rossol, 1994; Chatterjee et al., 1998). However, there is still no agreement on the mechanism of action of St. John's wort extracts or on the ingredient exerting the antidepressant effect. Originally, hypericin was thought to be the primary active compound and is the substance from which extracts from the plant were and continue to be standardized. However, recent data indicates that this may not be accurate and that hyperforin may be responsible for its antidepressant effect (Muller et al., 1997; Chatterjee et al., 1998). Additionally, Laakman et al. (1998) tested the efficacy of two preparations of St. John's wort that differed in hyperforin content. The data from this study suggests hyperforin content may be the critical active component for the antidepressant activity. Despite all these findings, the pharmacologically active components, which is responsible for its effects on mood is still not resolved. Additionally, because hyperforin is highly unstable when exposed to light and air, the content of hyperforin in the commercial preparations may vary considerably (Nahrstedt and Butterweck, 1997).

3.7.3 St. John's Wort for depression

In contrast to other herbal medicines, the efficacy of St. John's wort has been widely investigated in controlled trials (Linde et al., 1996). The majority of the studies are from Germany. In most of published trials, the patients had mild to moderate forms of depression.

Recent meta-analyses (Kim et al., 1999; Gaster and Holroyd, 2000; William et al., 2000; Whiskey et al., 2001; Linde and Mulrow, 2003) suggested that St John's wort is more effective than placebo for the short-treatment of mild to moderate depressive disorders. With the tricyclic antidepressants, amitriptyline and imipramine, they found St. John's wort to be of equivalent effectiveness, but with fewer side-effects.

However, several concerns apply to the clinical trails included in the meta-analyses comparing St. John's wort with placebo, or other antidepressants. For example, the combinability of the results of some of the St. John's wort trials is questionable. There was variability in St. John's wort preparations, heterogeneity of patient populations, inconsistent classification of depressive disorders, use of insufficient doses of active comparator drugs and variable trial durations (Wong et al., 1998; Linde and Mulrow, 2003). Additionally, Wong et al. (1998) raised concerns about the effectiveness of blinding given the difficulty in masking the taste of the extract.

3.8 Methodological issues in antidepressant trials

Antidepressant trials often include an inert placebo that may, in effect, unblind the studies because of the side effects of the active medication (Basoglu et al., 1997). Additionally, Schultz et al. (1995; 1996) have showed that insufficient randomisation and blinding and the exclusion of withdrawals from the analysis are able to inflate the apparent effect of an intervention.

Another problem encountered in antidepressant trials is the subjective nature of outcome assessment in depression. This could be a serious flaw because most studies rely primarily on potentially biased clinician-rated assessment (e.g., HAMD) rather than on patient-rated assessment (e.g., the Beck Depression Inventory). One meta-analysis (Lambert et al., 1986) showed that patient-rated measures led to significantly smaller effect sizes than clinician-rated measures.

Patients apparently tend to see less improvement compared to clinicians.

Also of concern is that some studies use a "washout" phase during which all participants are placed on placebo (Vorbach et al., 1994; 1997; Wheatley 1997). Those prospective participants who show improvement during the washout phase are eliminated from the pool of participants. This procedure may create a bias against the placebo condition because those "placebo responders" are excluded from the study. Thus, the actual placebo response rate may be seriously underestimated in those studies.

Murray (1989) has challenged the validity of HAMD because of the inclusion of many items that relate to anxiety or sleeping difficulties, as this may favour drugs with sedative properties.

The main challenge encountered when reviewers retrieve data from the primary studies for a systematic review is missing data. Often the investigators fail to report estimates of variance and only mean, test statistics such as F or t-values, or p-values are provided for observed changes in continuous outcome measures. Without estimates of standard deviation, these have to be imputed, or the results excluded from analysis.

3.9 Aims of Study

The aim for the second part of this project was to critically evaluate the evidence-base for St. John's Wort, perhaps the most widely used herbal medicine. The subobjectives identified on page 23 are reproduced here for ease of reference.

a. The characteristics of randomised controlled trials of St. John's Wort in the treatment of depression are described.

b. Methods of data imputation, approximation and transformation when reported data from trials are incomplete are discussed and applied.

c. Meta-analysis of trials of St. John's wort in depression is undertaken.

d. Sources of heterogeneity in randomised controlled trials of St. John's wort are critically evaluated.

CHAPTER 4

Characteristics of randomised controlled trials of St. John's wort for depression

4.1 Background and aim

Among the challenges to conduct systematic review is making a decision about whether the results of individual studies can be combined to arrive at an overall estimate of the treatment effect. Meta-analysis is a statistical technique that combines or integrates the results of several independent clinical trials considered by the analyst to be 'combinable'. One of the major criticisms directed towards meta-analysis of published studies is the pooling of heterogeneous studies, a practice which makes interpretation of any resultant pooled estimate of effect difficult and misleading. This problem is sometimes referred to as the mixing of apples and oranges. However, given that clinical studies, even those aiming to be replicates of each other, are never fully identical to each other, the basic problem is not per se, the pooling of results from studies with some differences in designs, but a loss of sight of factors which contribute to observed differences in effect. Without a clear view of potential sources of heterogeneity, the applicability of the results to a given practical scenario is difficult to determine.

The variation between studies which is often considered a weakness of a systematic review, however can be a considered a strength. The trials may differ in such factors as the intervention types, length of follow-up, characteristics of the study participants, severity of the disease and end points that were measured. If the results are consistent across many studies, despite variation in populations and methods, then the results may be considered robust and transferable. However, if the results are not consistent across studies then it is not appropriate to generalise the overall results. Any inconsistency between studies therefore presents a chance to explore the sources of variation and reach a deeper understanding of its causes and control (Thompson, 1994).

In this chapter we aimed to examine and describe the characteristics of randomised controlled trials of St. John's Wort for depression. We chose St. John's Wort because this herb is used by both practitioners of alternative medicine, and by conventionally trained physicians. There is therefore a potential clash in the perspectives of these two groups of carers, in relation to both management of patients and perception of the determinants of success in a clinical trial. Moreover, there is as yet no consensus as to whether St. John's Wort is effective in depression (Snow et al., 2000; Geddes and Butler, 2002).

4.2 Methods

4.2.1 Search strategy for the identification of studies

The following electronic bibliographic databases were searched: MEDLINE, PsycInfo, EMBASE and the Cochrane Central Register of Controlled Trials (CENTRAL). The electronic searches covered the period 1966 to the end of December 2002. There were no language restrictions.

A combination of terms was used to search the electronic databases (Table 4.1).

Table 4.1 Search terms used in the retrieval of articles from electronic databases

Field of focus	Search term
Study type	Clinical trial*, clinical, random*, controlled and double blind
Intervention-related	St. John's wort, Johanniskraut, hyperic*
Disease-related	Depress*, depressive-disorder*, mood*, neurotic*, and adjustment*

The abstracts for the list generated were then screened to exclude articles that were not randomised controlled trials. Bibliographies of articles obtained were also screened for further potentially relevant articles. Full copies were obtained for all potentially relevant papers and again screened to include only trials in which the primary focus was on treatment effects of St. John's wort for depression. Only double-blind randomised controlled trials that compared St. John's wort (as a single ingredient) with placebo or standard antidepressant treatments were selected.

If several papers by the same team of investigators with the same aim were available, these were considered as duplicates unless inclusion criteria for the papers were clearly different.

4.2.2 Data extraction

I extracted the data and my supervisor cross-validated them at random. We discussed and agreed on any ambiguity in the interpretation of the data. The following data were extracted: study identifications, diagnostic classification of depression, number of centre, number of patients, length of study, hypericum and hypericin dose per day, brand name, manufacturer and name and dosage of treatment comparator.

For outcome measure, two types of data were extracted (if available) from each study. The first type of data was dichotomous data, which was the outcome of recovered versus not recovered. The responder (recovered) was defined as the proportion of patients having either (a) 50% or greater improvement from baseline as assessed by HAMD rating scale or (b) a total score below a score of 10 or less on HAMD rating scale. The second, continuous data were scores using HAMD rating scale.

Data for 12 trials that were not published in English were extracted from the Cochrane Database of Systematic Review [St. John's Wort for Depression; by Linde and Mulrow (2003)].

4.3 Results

Of the 108 citations originally identified in our electronic search, 57 were potentially relevant, the abstracts of which were retrieved and reviewed. Application of the exclusion criteria excluded 27 trials (Appendix 6), which brought the total of included studies to 30 (Appendix 5) consisting exclusively of randomised controlled studies. Eighteen of the selected trials were published in English and the remaining 12 trials in German (Appendix 5). Of these 30 trials, three trials have more than two treatment arms [one trial of St. John's with two different doses of hyperforin (Laakmann et al., 1998) and the other two trials (Hypericum Depression Trial study Group, 2002; Philipp et al., 1999) have more than one control group (an established antidepressant and a placebo)].

Sixteen of these trials were placebo-controlled and 11 trials compared St. John's Wort with other antidepressants.

The studies differ on a number of dimensions:

4.3.1. Demographic of patients, study size and setting of trials

The mean age of patients from the trials ranged from 40 to 68.8 years (Table 4.2). All but one of the trials had more female than male. In one trial (Harrer et al., 1999) the percentage of female participants was close to 90%.

The largest randomized controlled trials of St. John's wort is by Lecrubier et al. (2002) with 375 participants followed by the trial of Woelk (2000) with a sample size of 324. However, seven trials have a sample size of 50 and less (Table 4.2).

The majority of the trials were conducted in Europe (mainly Germany). However, three recent trials were conducted in the USA. There are eight single-center studies. The remaining 22 are multicentre studies involving between 2-40 centers (Table 4.2).

4.3.2 Diagnostic classifications and duration of trials

Diagnostically, the groups of patients in the studies were heterogeneous. Four different diagnostic criteria (ICD-9, ICD-10, DSM-IIIR, and DSM-IV) were used (Table 4.3).

In addition to diagnostic criteria, studies differed widely in terms of the baseline HAMD scores at entry. The scores ranged from 13 to 24. With the exception of one study (Vorbach et al., 1997), all other trials reported the inclusion of patients with mild to moderately severe depression. However, even though Schmidt (1989) and Schlich (1987) reported patients in their trials were suffering from mild to moderate depression, the baseline values of these trials were suggestive of severe depression (Snow et al., 2000).

Sixteen trials used the 17-items version of the HAMD rating scale as the primary outcome measurement. Two trials used the 21-items version, while for the remaining trials, the version of the scale used was either not reported or unknown (Table 4.3).

The duration of trials vary from 4 to 8 weeks with one longest trial lasting for 12 weeks. However, the majority of the trials that used the ICD 9 diagnostic classification were of 4 weeks duration (Table 4.2)

4.3.3 Different types of St. John's wort preparations

Several different preparations (with various dosage forms) from ten different manufacturers were tested. Both solid and liquid dosage forms were employed in the studies. Daily doses of extract and amount of total hypericin, which is the reference substance for pharmaceutical standardisation varied considerably (between 300 and 1080 mg and 0.4 and 2.7 mg, respectively) among trials (Table 4.4).

4.3.4 Outcome measures (continuous data)

The 30 identified studies that used HAMD instrument reported continuous outcomes in a variety of ways, including means at endpoint, median, and mean change-from-baseline. Additionally the variability was also reported using several different statistics namely standard deviation, standard error and confidence interval.

The outcome measures were reported as final HAMD in 16 studies, one of which did not give any estimate of variability (Harrer et al., 1999). Change-from-baseline was reported in seven studies (Hypericum Trial Study Group, 2002; Lecrubier et al., 2002; Phillip et al., 1999; Schrader et al., 1998; Schrader 2000; Shelton et al., 2001; Woelk et al., 2000). Both final score and change-from-baseline was reported in four studies (Brenner et al., 2000; Kalb et al., 2001; Laakmann et al., 1998; Van Gurp et al., 2002) and Wheatley (1997) reported median score and two other studies (Halama, 1991; Schlich et al., 1987) reported only binary outcomes.

4.3.5 Outcome Measures (binary data)

Except for two trials (Shelton et al., 2001; Hypericum Study Group, 2002) all the other studies including placebo and antidepressant comparison studies, report superiority of St. John's wort when compared to placebo and equivalence to antidepressants such as imipramine, fluoxetine, sertraline and maprotiline (Table 4.5).

4.4. Discussion

The results of our study are in agreement with Linde and Mulrow (2003) who noted that the randomised controlled trials of St. John's wort for depression were heterogeneous. Heterogeneity of studies is unavoidable. The question is not whether it is present but whether its extent seriously undermines the conclusions being drawn. Thus, consideration of heterogeneity has important implications for the design and interpretation of meta-analyses, even in apparently focused clinical areas.

The diversity of the trials was particularly marked with respect to diagnostic classification of depression and the severity of the depression at entry. Eight trials used ICD-9 classifications. The diagnoses in these trials would probably correspond to the DSM-IV categories of adjustment disorder with depressed mood (code: 309.0) or acute stress disorder (code: 308.3) rather than major depression (code 296.XX). As such this brings into question the wisdom of pooling data from the St. John's wort trials assessing efficacy in different types of depression.

Additionally, the HAMD scores at entry for the trials varied from 13 to 25; thus the severity of depression studied in the trials ranged from mild to moderately severe according to established cut-off scores (Snow et al., 2000). Therefore, patients who were more mildly depressed might have a different outcome to treatments compared to those who were more severely depressed. Laakmann et al. (1998) has suggested that treatment with St. John's wort was more efficacious for the more severely depressed patients. Thus the severity of the patient's depression and its relationship to treatment efficacy has important clinical implications. According to the evidence presented by Linde and Mulrow (2003) and in other reviews (William et al., 2000; Whiskey et al., 2001), St. John's Wort seems to be effective for almost any type of depression. This may be correct, but further work needs to address this question and should attempt to determine which types of depression are optimally effective to treatment with St. John's wort and which are not.

The optimum adult dose of St. John's wort for treating depression, based on the included studies, appears to be 300 mg of plant extract orally three times daily. However, doses used varied considerably among studies. Currently, manufacturers of herbal medicines are not required to produce a product that meets set standards for uniformity and consistency. As such this raises concerns regarding variability of St. John's wort from one manufacturer to the next. Concerns include composition, quality, dosage, purity, and potency. Furthermore, the amount of active substances might vary depending on factors such as the extraction process, season, and plant parts used. Wide variations also have been found in the concentration of the active ingredient (believed to be hypericin) in different preparations of St. John's wort, despite labelled doses (Wong et al., 1998; Busse, 2000). Thus, a fundamental problem in the randomised controlled trials of St. John's wort is whether different products, extracts, or even different lots of the same extract are comparable and equivalent. Pooling studies that use different St. John's wort preparations in a quantitative meta- analysis can be misleading.

For continuous data, the trials were not homogenous in the way treatment effects are reported. Treatment response was either reported as final HAMD score or change-from-baseline values. This would pose practical problems for a meta-analyst.

For binary data, the outcome was defined in terms of a 50% change on HAMD scale or a cut-off score of 10 on the HAMD scale. Although the HAMD scale was used for all the trials, different versions with different scores were employed. The use of several versions of the HAMD scale with different numbers of items causes difficulties in interpretation of results from different trials (Hedlund and Vieweg, 1979). This heterogeneity was compounded by incomplete reporting with respect to the version of the scale used.

Placebo response varied from trial to trial. It was high and close to 50% for two trials. Placebo response in clinical trials of antidepressants is known to range from 30 to 60% (Gavin, 2001). Because of the high and variable placebo response rate, in the absence of a placebo group (for comparative studies with antidepressants) no firm conclusion can be drawn about the efficacy of St. John's wort, even when the response rates for both hypericum and antidepressants groups are comparable. There is a chance that both treatments would be shown to be ineffective had the placebo arm been included. Given that in most trials the doses of standard antidepressant used as comparator were below the normal doses (imipramine 50 or 75mg daily; maprotiline 75mg daily), this doubt about the efficacy of St. John's wort is increased.

4.5 Conclusion

Published randomised controlled trials addressing whether St. John's wort is more effective than placebo or as effective as standard antidepressants in the treatment of depression were heterogeneous with respect to many influential variables. Therefore unless these are taken into account, estimates of effect St. John's wort can be seriously misleading.

TABLE 4.2 Characteristics of study (demographic of patients and trials setting)

Preparation of St. John's Wort (Manufacturer)	Trials	Comparator	No. of patients	Mean Age (Years)	Female (%)	Setting (No. of Centres)	Country of Trials	Daily Dose of Extract (mg)
LI 160 Extract Tablet (Lichtwer Pharma GmbH, Germany)	Brenner et al. (2000)	Sertraline	30	45.5	63.3	Single Centre	USA	600-900
	Halama (1991)	Placebo	50	47.0	64.0	Single Centre	Germany	900
	Hängsen et al. (1996)	Placebo	108	52.0	61.0	Multicentre (17)	Germany	900
	Harrer et al. (1994)	Maprotiline	102	48.0	71.6	Multicentre (6)	Germany	900
	Hübner et al. (1994)	Placebo	40	51.0	56.4	Single Centre	Germany	900
	Hypericum Depression Trial Study Group (2002)	Placebo and Sertraline	340	42.3	65.9	Multicentre (12)	USA	900-1800
	Lehrl et al. (1993)	Placebo	50	49.0	82.0	Single Centre	Germany	900
	Schmidt & Sommer (1993)	Placebo	65	44.0	76.9	Multicentre (3)	Germany	900
	Shelton et al. (2001)	Placebo	200	42.4	64.0	Multicentre (11)	USA	900-1200
	Sommer & Harrer (1994)	Placebo	105	48.0	72.4	Multicentre (3)	Germany	900
	Vorbach et al. (1994)	Imipramine	135	53.0	47.4	Multicentre (20)	Germany	900
	Vorbach et al. (1997)	Imipramine	209	49.0	73.7	Multicentre (20)	Germany	1800
	Wheatley (1997)	Amitriptyline	165	40.0	76.4	Multicentre (19)	UK	900
Psychotonin® M Liquid Extract (Steigerwald, Darmstadt, Germany)	Harrer et al. (1991)	Placebo	120	49.0	59.2	Multicentre (6)	Germany	500
	Quandt et al. (1993)	Placebo	88	43.0	65.9	Multicentre (4)	Germany	500
	Schlich et al. (1987)	Placebo	46	42.0	63.0	Single Centre	Germany	350
	Schmidt et al. (1989)	Placebo	40	47	45.0	Multicentre (2)	Germany	500
Psychotonin® M Capsule (Steigerwald, Darmstadt, Germany)	Witte et al. (1995)	Placebo	97	43.0	66.0	Multicentre (5)	Germany	200-240

Preparation of St. John's Wort (Manufacturer)	Trials	Comparator	No. of patients	Mean Age (Years)	Female (%)	Setting (No. of Centres)	Country of Trials	Daily Dose of Extract (mg)
WS 5570 (Dr. Willmar Schwabe Pharmaceutical, Germany)	Lecrubier et al. (2002)	Placebo	375	40.7	76.5	Multicenter (26)	France	900
WS 5572 and WS 5573 Extract Tablet (Dr. Willmar Schwabe, Germany)	Laakmann et al. (1998)	Placebo	147	49.0	79.6	Multicentre (11)	Germany	900
WS 5572 (Dr. Willmar Schwabe, Germany)	Kalb et al. (2001)	Placebo	72	48.5	66.7	Multicenter (11)	Germany	900
Neuroplant® Capsule (Dr. Willmar Schwabe, Germany)	Reh et al. (1992)	Placebo	50	48.0	78.0	Single Centre	Germany	500
ZE117 Extract Tablet (Zeller AG, Switzerland)	Schrader et al. (1998) Schrader et al. (2000)	Placebo Fluoxetine	165 238	- 46.5	65.5 65	Multicentre (16) Multicenter (7)	Germany Germany	500 500
Remotiv® Tablet (Bayer Vital, Leverkusen, Germany)	Woelk (2000)	Imipramine	324	46.0	71.3	Multicentre (40)	Germany	500
Esbericum® Capsule (Schaper & Brümmer, Germany)	Bergmann et al. (1993)	Amitriptylin	80	55.0	66.3	Single Centre	Germany	??

Preparation of St. John's Wort (Manufacturer)	Trials	Comparator	No. of patients	Mean Age (Years)	Female (%)	Setting (No. of Centres)	Country of Trials	Daily Dose of Extract (mg)
Dysto-lux® Tablet (Dr. Loges + Co GmbH, Germany)	Harrer et al. (1999)	Fluoxetine	149	68.8	86.6	Multicentre (17)	Germany	800
Calmigen® Tablet (SanoPharm Skelstedet, Denmark)	Behnke et al. (2002)	Fluoxetine	69	49.7	68.1	Multicentre (?)	Denmark	300
STEI 300 Extract Capsule (Steiner Arzneimittel, Germany)	Philipp et al. (1999)	Impipramine and Placebo arm	263	47.0	74.9	Multicentre (18)	Germany	1050
Hypericum Extract imported from Germany	Van Gurp et al. (2002)	Sertraline	87	40.0	59.8	Single Center	Canada	900

Table 4.3 Characteristics of study: Diagnostic classifications and assessment durations

Diagnostic Classification System	Study	Classification of depressive disorders	Classification Codes	Baseline HAMD scores	HAMD items	Endpoint (weeks)
DSM III-R	Hängsen et al. (1996)	Major depression	Not given	>15	17	4
	Vorbach et al. (1994)	Major depression with single episode Major depression with recurrent episodes Depressive neurosis Adjustment disorder	296.2 296.3 300.4 309.0	NR	17	6
DSM IV	Brenner et al. (2000)	Major depressive disorder single episode Major depressive disorder recurrent Dysthymic disorder Adjustment disorder with depressed mood Depressive disorder not otherwise specified	296.21 296.31 300.4 309.0 311	≥ 17	17	7
	Hypericum Depresion Trial Study Group (2002)	Major depressive disorder	Not reported	≥ 20	17	8
	Kalb et al. (2001)	Major depressive disorder single episode Major depressive disorder single episode (moderate) Major depressive disorder recurrent Major depressive disorder recurrent (moderate)	296.21 296.22 296.31 296.32	≥ 16	17	6
	Laakmann et al. (1998)	Major depressive disorder single episode Major depressive disorder single episode, (moderate) Major depressive disorder recurrent Major depressive disorder recurrent (moderate)	296.21 296.22 296.31 296.32	≥ 17	17	6

Diagnostic Classification System	Study	Classification of depressive disorders	Classification Codes	Baseline HAMD scores	HAMD items	Endpoint (weeks)
	Lecrubier et al. (2002)	Major depressive disorder single episode Major depressive disorder single episode (moderate) Major depressive disorder recurrent Major depressive disorder recurrent (moderate)	296.21 296.22 296.31 296.32	18-25	17	6
	Shelton et al. (2001)	Major depression single episode Major depression recurrent		≥ 20	17	8
	Van Gurp (2002)	Major depression	NR	≥ 16	17	12
	Wheatley (1997)	Major depression	NR	17-24	17	6
ICD-10	Behnke et al. (2002)	Mild depressive episode Moderate depressive episode	F32.0 F32.1	16-24	17	6
	Bergmann et al. (1993)	Mild depressive episode Moderate depressive episode Recurrent depressive disorder, current episode (mild) Recurrent depressive disorder, current episode (moderate)	F32.0 F32.1 F33.0 F33.1	NR	NR	6
	Harrer et al. (1994)	Moderate depressive episode	F32.1	≥ 16	17	4
	Harrer et al. (1999)	Mild depressive episode Moderate depressive episode	F32.0 F32.1	NR	17	6
	Philipp et al. (1999)	Moderate depressive episode Recurrent depressive disorder, current episode (moderate)	F32.1 F33.1	≥ 18	17	6,8

Diagnostic Classification System	Study	Classification of depressive disorders	Classification Codes	Baseline HAMD scores	HAMD items	Endpoint (weeks)
	Schrader et al. (1998)	Mild depressive episode Moderate depressive episode	F32.0 F32.1	16-24	21	6
	Schrader et al. (2000)	Mild depressive episode Moderate depressive episode	F32.0 F32.1	16-24	21	6
	Vorbach et al. (1997)	Recurrent depressive disorder, current episode severe without psychotic symptoms	F33.2	NR	17	7
	Witte et al. (1995)	Moderate depressive episode	F32.1	≥ 16	NR	6
	Woelk (2000)	Mild depressive episode Moderate depressive episode Recurrent depressive disorder, current episode (mild) Recurrent depressive disorder, current episode (moderate)	F32.0 F32.1 F33.0 F33.1	≥ 18	17	6
ICD-9	Halama (1991)	Neurotic depression Adjustment disorder	300.4 309.0	16-20	NR	4
	Harrer et al. (1991)	Neurotic depression Adjustment disorder	300.4 309.0	NR	NR	6
	Hübner et al. (1994)	Neurotic depression Adjustment disorder	300.4 309.0	*	NR	4
	Lehrl et al. (1993)	Neurotic depression Adjustment disorder	300.4 309.0	16-20	NA	4

Diagnostic Classification System	Study	Classification of depressive disorders	Classification Codes	Baseline HAMD scores	HAMD items	Endpoint (weeks)
	Quandt et al. (1993)	Neurotic depression	300.4	>15	NA	4
	Reh et al. (1992)	Neurotic depression Adjustment disorder	300.4 300.9	NR	NA	8
	Schmidt et al. (1989)	Mild to moderately severe depressive disorder	NR	>15	NA	4
	Schmidt & Sommer (1993)	Neurotic depression Adjustment disorder	300.4 309.0	16-20	NA	6
	Sommer & Harrer, (1994)	Neurotic depression Adjustment disorder	300.4 300.9	≤ 20	NA	4
Not Reported	Schlich et al. (1987)	Mild to moderately severe depressive disorder	Not given	>15	NA	4

Notes:
* - HAMD score at baseline for inclusion was not given. However, at entry into the study, the mean HAMD score for the included patients was < 13.0.

NA - Not Available
NR – Not Reported

Table 4.4 Characteristics of study: Types of St. John's Wort preparations and daily dose

Product Name (Name of Extract)	Manufacturer	Study	Dosage Form	Labelled amount of Extract (mg)	Hypericin (%)	Total Daily Dose of Extract (mg)	Total daily dose of hypericin (mg)
Aristo® 350 (STEI 300)	Steiner Arzneimittel, Germany	Phillip et al. (1999)	Capsule	350	0.2-0.3	1050	2.1-3.15
Calmigen®	SanoPharm Skelstedet, Denmark	Behnke et al. (2002)	Coated tablet	150	0.3-0.33	300	0.9-0.99
Dysto-Lux® (LoHyp-57)	Dr. Loges+Co GmbH, Germany	Harrer et al. (1999)	Coated Tablet	200	NA	800	NA
Esbericum®	Schaper & Brümmer, Germany	Bergman et al. (1993)	Capsule	NA	NA	NA	0.75
Jarsin® (LI 160)	Lichtwer Pharma GmbH, Germany	Lehrl et al. (1993)	Sugar Coated Tablet	300	0.12	900	1.08
		Halama (1991)	Sugar Coated Tablet	300	0.12	900	1.08
		Schmidt & Sommer (1993)	Sugar Coated Tablet	300	0.12	900	1.08
Jarsin® 300 (LI 160)	Lichtwer Pharma GmbH, Germany	Brenner et al. (2000)	Sugar Coated Tablet	300	0.3	600-900	1.8-2.7
		Hangsen et al. (1994)	Sugar Coated Tablet	300	0.3	900	2.7
		Hansgen (1996)	Sugar Coated Tablet	300	0.3	900	2.7
		Harrer et al. (1994)	Sugar Coated Tablet	300	0.3	900	2.7
		Hypericum Depression Trial Study Group (2002)	Sugar Coated Tablet	300	0.3	900-1800	3-5
		Shelton et al. (2001)	Sugar Coated Tablet	300	0.3	900-1200	2.7-3.6
		Sommer & Harrer (1994)	Sugar Coated Tablet	300	0.3	900	2.7

Product Name (Name of Extract)	Manufacturer	Study	Dosage Form	Labelled amount of Extract (mg)	Hypericin (%)	Total Daily Dose of Extract (mg)	Total daily dose of hypericin (mg)
		Vorbach et al. (1994)	Sugar Coated Tablet	300	0.3	900	2.7
		Vorbach et al. (1997)	Sugar Coated Tablet	300	0.3	up to 1800	up to 5.4
		Wheatley (1997)	Sugar Coated Tablet	300	0.3	900	2.7
Neuroplant®	Dr. Willmar Schwabe Pharmaceuticals, Germany	Reh et al. (1992)	Capsule	125	0.2	500	0.75
Psychotonin® M forte	Steigerwald Darmstadt, Germany	Witte et al. (1995)	Capsule	200-240	0.5	200-240	1.0-1.2
Psychotonin M®	Steigerwald Darmstadt, Germany	Harrer et al. (1991)	Tincture	5.56mg/ drop	NA	500	0.75
		Quandt et al. (1993)	Tincture	5.56mg/ drop	NA	500	0.75
		Schlich 1987	Tincture	5.83mg/ drop	NA	350	0.5
		Schmidt et al. (1989)	Tincture	5.56mg/ drop	NA	500	0.75
Remotiv® (ZE 117)	Bayer Vital, Leverkusen, Germany	Woelk (2000)	Film-coated tablet	250	0.2	500	1
-	Dr.Willmar Schwabe Pharmaceuticals, Germany	Laakmann et al. (1998)	Tablet	300	0.5%*	900	4.5*
(WS 5573)				300	5.0% *	900	45.0*
(WS 5572)		Kalb et al. (2001)	Film-coated tablet	300	1.5%*	900	
WS 5570	Dr.Willmar Schwabe Pharmaceuticals, Germany	Lecrubier et al. (2002)	Film-coated tablet	300	0.12-0.28%	900	1.08-2.52

Product Name (Name of Extract)	Manufacturer	Study	Dosage Form	Labelled amount of Extract (mg)	Hypericin (%)	Total Daily Dose of Extract (mg)	Total daily dose of hypericin (mg)
-	Zeller AG, Switzerland	Schrader et al. (1998)	Film-coated tablet	250	0.2	500	1
(ZE 117)		Schrader et al. (2000)	Film-coated tablet	250	0.2	500	1
Not Available	Not Available	Van Gurp et al. (2002)	Capsule	300	NA	900	NA

Notes:

*	Hyperforin

NA - Not Available

Table 4.5 Dichotomous outcomes (response rate)

Study	Comparator (daily dose mg)	Sample Size N	Responders (St. John's wort) n/N	Responders (St. John's wort) %	Responders (Comparator) n/N	Responder (Comparator) %	Results
Behnke (2002)	Fluoxetine (40)	70	16/29	55.17	25/32	78.13	Comparable efficacy to fluoxetine
Bergmann et al. (1993)	Amitriptyline (90)	80	32/40	80.00	28/40	70.00	Favoured St. John's wort
Brenner et al. (2000)	Sertraline (75)	30	7/15	46.67	6/15	40.00	Comparable efficacy to sertraline
Harrer et al. (1994)	Maptrotiline (75)	102	27/51	52.94	28/51	54.90	Comparable efficacy to maprotiline
Harrer et al. (1999)	Fluoxetine (20)	149	50/70	71.43	57/79	72.15	Comparable efficacy to fluoxetine
Schrader (2000)	Fluoxetine (20)	240	75/125	60.00	45/113	39.82	Comparable efficacy to fluoxetine
van Gurp (2002)	Sertraline (50-100)	87	NA	NA	NA	NA	Comparable efficacy to sertraline
Vorbach et al. (1994)	Imipramine (75)	135	42/67	62.69	37/68	54.41	Comparable efficacy to imipramine
Vorbach et al. (1997)	Imipramine (150)	209	38/107	35.51	42/102	41.18	Comparable efficacy to imipramine
Wheatley (1997)	Amitriptyline (75)	165	50/87	57.47	57/78	73.08	Amitriptyline slightly better
Woelk (2000)	Imipramine (150)	324	68/157	43.31	67/167	40.12	Comparable efficacy to imipramine
Halama (1991)	Placebo	50	10/25	40.00	0/25	0.00	Favoured St. John's wort
Hängsen et al. (1996)	Placebo	108	35/53	66.04	12/54	22.22	Favoured St. John's wort
Harrer et al (1991)	Placebo	120	NA	NA	NA	NA	Favoured St. John's wort
Hübner et al. (1994)	Placebo	40	14/20	70.00	9/20	45.00	Favoured St. John's wort
Kalb (2001)	Placebo	72	23/37	62.16	15/35	42.86	Favoured St. John's wort

Study	Comparator (daily dose mg)	Sample Size N	Responders (St. John's wort) n/N	Responders (St. John's wort) %	Responders (Comparator) n/N	Responder (Comparator) %	Results
Lecrubier (2002)	Placebo	375	98/186	52.69	80/189	42.33	Favoured St. John's wort
Lehrl et al. (1993)	Placebo	50	4/25	16.00	2/25	8.00	Favoured St. John's wort
Quandt et al. (1993)	Placebo	88	29/44	65.91	4/33	6.82	Favoured St. John's wort
Reh et al. (1992)	Placebo	50	20/25	80.00	11/25	44.00	Favoured St. John's wort
Schlich et al. (1987)	Placebo	46	15/25	60.00	3/24	12.50	Favoured St. John's wort
Schmidt and Sommer (1993)	Placebo	65	20/32	62.50	6/33	18.18	Favoured St. John's wort
Schmidt et al. (1989)	Placebo	40	10/20	50.00	4/20	20.00	Favoured St. John's wort
Schrader et al. (1998)	Placebo	162	45/81	55.56	12/81	14.81	Favoured St. John's wort
Shelton et al. (2001)	Placebo	200	26/98	26.53	19/102	18.63	St. John's wort not effective
Sommer & Harrer, (1994)	Placebo	105	28/50	56.00	13/55	23.64	Favoured St. John's wort
Witte et al. (1995)	Placebo	97	34/48	70.83	25/49	51.02	Favoured St. John's wort
St. John's wort Depression Trial Study Group (2002)*	Sertraline (50-100)	340	16/113	14.16	26/109	23.85	St. John's wort not effective
	Placebo				13/116	11.21	
Laakmann et al. (1998)*	Placebo	147	19/49 24/49	38.78 48.98	16/49	32.65	Effectiveness of St. John's wort depends on hyperforin content
Philipp et al. (1999)*	Placebo	263	67/100	67.00	22/46	47.83	More effective than placebo
	Imipramine				66/105	62.86	As effective as Imipramine

* Trials with 3 treatment arms

Responders are based on the criteria of ≥ 50% reduction in HAMD score from baseline or a total score of ≤ 10 at the end of treatment.

CHAPTER 5

Methods of data imputation, approximation and transformation when reported data are missing, incomplete or in different forms

5.1 Introduction and aim

Meta-analyses are usually dependent on summary data obtained from published reports of clinical trials to provide an estimate of treatment effect. Estimating effect of treatment using continuous data usually involves comparing means of treatment and control groups. Estimates of variability are also required. However, reports of randomised controlled trials (RCTs) are often incomplete and data are often reported in alternative formats. For example, in randomised controlled trials of St. John's wort for depression, one of the problems faced by reviewers is the failure of several primary trials to report an estimate of variance. Unless this estimate can be imputed, the results of the trials cannot be included in any meta-analysis. It is also common to find studies using different measures of variability, such as standard deviation, standard error and confidence interval. If the standard deviation is not reported directly, it may be calculated from the standard error or confidence interval provided the sample sizes are given.

Another practical problem encountered when performing a meta-analysis of continuous data in St. John's wort trials is the reporting of treatment effects in different forms. For example, treatment response may be reported as final HAMD score or change from-baseline values.

Thus, to take account of as much of the published data as possible when undertaking a meta-analysis of those trials, there is a need to standardise the observed effects and their associated variability. This involves data transformation, approximation and imputation. In this chapter, we describe the methods for data transformation and imputation, which may be appropriate to estimate treatment effects so that pooled analyses can be conducted in an assessment of randomised controlled trials of St John's wort for depression.

5.2 Methods and results

5.2.1 Identification, selection of articles and data extraction

Of the 30 trials, which met the inclusion criteria (chapter 4), two trials (Halama, 1991; Schlich et al., 1987) reported the outcome as a binary data. As such no data manipulation can be done to enable the results to be analysed as continuous outcomes.

When available we extracted values on final means, change-from-baseline, and measure of variability or other statistics both before (baseline) and after treatment. Where possible data were extracted based on an intention-to-treat analysis.

5.2.2 Presentations of treatment outcomes

The 28 selected studies that reported outcome as a continuous data presented the results in a variety of ways, including mean at endpoint, median, and mean change-from-baseline. Additionally the measure of variability was also reported using several different statistics such as SD, SEM and CI.

The outcome measures were reported as final HAMD in16 studies, one of which did not report measure of variability (standard deviation). Change-from-baseline is reported in seven studies; both final score and change-from-baseline in four studies; and one study reported median score (Table 5.1)

Table 5.1: Presentation of treatment outcomes

Presentation of outcomes	Number of trials	Trials
Final HAMD score (mean and SD)	15	Behnke et al., 2002; Bergman et al., 1993; Hängsen et al., 1996; Harrer et al., 1994; Harrer & Sommer 1994; Hübner et al., 1994; Lehrl et al., 1993; Quandt et al., 1993; Reh et al., 1992; Schmidt et al.,1989; Schmidt & Sommer, 1993; Sommer & Harrer, 1994; Vorbach et al., 1994; Vorbach et al., 1997; Witte et al., 1995
Final HAMD score (mean without SD)	1	Harrer et al., 1999
Final HAMD score (median and range)	1	Wheatley, 1997
Change-from-baseline (mean and SD)	3	Lecrubier et al., 2002; Philipp et al., 1999; Shelton et al., 2001
Change-from-baseline (mean and SEM)	1	Hypericum Trial Study Group, 2002
Change-from-baseline (mean and CI)	2	Schrader et al., 1998; Schrader, 2000
Change-from-baseline (mean without SD)	1	Woelk et al., 2000
Both Final HAMD score and change-from-baseline (mean with SD)	4	Brenner et al., 2000; Kalb et al., 2001; Laakman et al., 1998; Van Gurp et al., 2002
Only dichotomous outcome	2	Halama, 1991; Schlich et al., 1987

5.2.3 Data conversion (confidence interval to standard deviation)

In two papers, (Schrader 1998; Schrader 2000), standard deviation of the change-from-baseline is not reported directly. Confidence interval (CI) was converted to standard deviation (SD) by using the following formula:

$\mathrm{CI} = \mathrm{Mean} \pm [Z_{1-\alpha/2}] \times \mathrm{SE}$

Where $\mathrm{SE} = \mathrm{SD}/\sqrt{n}$

At an α level of 0.05, $[Z_{1-\alpha/2}] = 1.96$

Therefore the 95% confidence interval is given by

$\mathrm{CI} = \mathrm{Mean} \pm 1.96 \times \mathrm{SE}$

$\mathrm{SD} = \sqrt{n}\,(\mathrm{CI})/\,(2 \times 1.96)$

5.2.4. Imputation of missing data (standard deviation)

Harrer et al. (1999) reported mean values at baseline and end of treatment without SD. Unless these estimates of standard deviations can be imputed, these results cannot be pooled and included in the meta-analyses. We estimated the missing standard deviation (for baseline) by pooling standard deviation from all other studies reporting this value (SD) for baseline using the method described by Follmann et al. (1995).

$$\text{Pooled variance} = \frac{[(n_1-1)s_1^2 + (n_2-1)s_2^2 + \ldots (n_k-1)s_k^2\ ]}{n_1 + n_2 + \ldots n_k}$$

where 1, 2,…k refers to the different trials

and n_1, n_2,…n_k refers to the sample size of the different trials.

Pooled standard deviation = √ (pooled variance)

In this calculation it is assumed that there exists a single underlying standard deviation of which the pooled standard deviation is a better estimate than the individual standard deviations of s_1, s_2…s_k.

For the missing SD at endpoint, we pooled SD from all other studies with the same comparator (fluoxetine) using the above formula.

5.2.5 Transformation to change-from-baseline

Sixteen trials including one trial with imputed SD reported treatment effect as final HAMD values (Table 5.1). The mean change-from-baseline and the SD of the change for each trial were derived as illustrated by the following worked example.

a. Calculation of mean change-from-baseline

Worked example: (Values from Kalb et al., 2001)

	Baseline HAMD score		Final HAMD Score	
	Mean	SD	Mean	SD
Hypericum (n=37)	19.7	3.4	8.9	4.3
Control (n=35)	20.1	2.6	14.4	6.8

Mean difference = mean at endpoint minus mean at baseline

In the Hypericum group the change-from-baseline is 19.7 – 8.9 = 10.8

In the Control group the change-from-baseline is 20.1 – 14.4 = 5.7

b. Calculation of standard deviation of the change

Standard error (SE) difference = $\surd\ [SD_1^2/n_1 + SD_2^2/n_2]$

Where:

SD_1 is the SD at baseline
n_1 is the sample size at baseline
SD_2 is the SD at end of treatment
n_2 is the sample size at end of treatment

To calculate the SD difference from the SE difference:

SE= SD/ $\sqrt{n}$

SD difference = SE difference x $\sqrt{n}$

Using the above formula, the SD difference in the hypericum group is:

SE difference Hypericum group = $\sqrt{[3.4^2/37 + 4.3^2/37]}$

= 0.81

SD difference Hypericum group = 0.81 x $\sqrt{37}$

= 4.98 (Kalb et al., 2001 reported a SD of 5.0)

The same process is used to calculate the SD difference in the control group.

This approach allows the pooling of variance from baseline and end of treatment values. However, it does not take into consideration that repeated measurements (at baseline and after treatment) made on the same participants tend to be correlated. Therefore the variance estimate for the difference is conservative.

Wheatley (1997) indicated that the outcome might have a skewed distribution and thus reported HAMD median and range. As there are currently no reliable methods available to approximate means from median, this type of data could not be converted to change-from-baseline values.

5.2.6 Summary of data approximation, imputation and transformation

Table 5.2 summarises how data were approximated, imputed, converted or transformed into a standard common measure (change-from-baseline and the SD of the change). Since all the included studies now measured the outcome on the same scale, the Weighted Mean Difference (WMD) method can be used (Der Simonian and Laird, 1986), to obtain a summary effect and its confidence interval. A fixed effect and random effects method may be used depending on the test of heterogeneity. If heterogeneity is absent, the fixed effect model is used to report results; otherwise the random effects model is used (Der Simonian and Laird, 1986).

Table 5. 2 Summary of data handling

Study	Mean Final HAMD	SD	Change - from - baseline	SD	Data handling
Bergman et al. (1993)	✓	✓	–	–	convert final HAMD to change-from-baseline
Wheatley (1997)	HAMD median	range	–	–	could not convert data to change-from-baseline
Behnke et al. (2002)	✓	✓	–	–	convert final HAMD to change-from-baseline
Harrer et al. (1999)	✓	–	–	–	impute missing SD→ convert final HAMD to change-from-baseline
Schrader et al. (2000)	✓	CI	✓	CI	convert CI to SD for change-from-baseline

Study	Mean Final HAMD	SD	Change - from - baseline	SD	Data handling
Phillip et al. (1999)	Graph	Graph (SEM)	✓	–	impute SD for change-from-baseline
Vorbach et al. (1994)	✓	✓	–	–	convert final HAMD to change-from-baseline
Vorbach et al. (1997)	✓	✓	–	–	convert final HAMD to change-from-baseline
Woelk (2000)	–	–	✓	–	impute SD for change-from-baseline
Harrer et al. (1994)	✓	✓	–	–	convert final HAMD to change-from-baseline
Brenner et al. (2000)	✓	✓	✓	✓	data directly useable
Hypericum Depression Trial Study Group (2002)	–	–	✓	SEM	convert SEM to SD
van Gurp (2002)	✓	✓	✓	✓	data directly useable
Halama (1991)	–	–	–	–	could not analyse as continuous outcome
Hänsgen et al. (1996)	✓	✓	–	–	convert final HAMD to change-from-baseline
Harrer & Sommer (1994)	✓	✓	–	–	convert final HAMD to change-from-baseline
Hübner et al. (1994)	✓	✓	–	–	convert final HAMD to change-from-baseline
Kalb et al. (2001)	✓	✓	✓	✓	data directly useable
Laakmann et al. (1998)	✓	✓	✓	✓	data directly useable

Study	Mean Final HAMD	SD	Change - from - baseline	SD	Data handling
Lecrubier et al. (2002)	–	–	✓	✓	data directly useable
Lehrl et al. (1993)	✓	✓	–	–	convert final HAMD to change-from-baseline
Quandt et al. (1993)	✓	✓	–	–	convert final HAMD to change-from-baseline
Reh et al. (1992)	✓	✓	–	–	convert final HAMD to change-from-baseline
Schlich et al. (1987)	–	–	–	–	could not analyse as continuous outcomes
Schmidt et al. (1989)	✓	✓	–	–	convert final HAMD to change-from-baseline
Schmidt & Sommer (1993)	✓	✓	–	–	convert final HAMD to change-from-baseline
Schrader et al. (1998)	–	–	✓	CI	convert CI to SD
Shelton et al.(2001)	Graph	Graph	✓	✓	data directly useable
Sommer & Harrer (1994)	✓	✓	–	–	convert final HAMD to change-from-baseline
Witte et al. (1995)	✓	✓	–	–	convert final HAMD to change-from-baseline

5.3 Discussion and conclusions

Meta-analysis is a useful statistical procedure that combines the results of several independent studies considered to be combinable. When the outcome is a continuous data, both treatment effect and variance are required for each trial. Often only partial information is available in the published studies. Systematic biases may occur if results of continuous data are excluded in meta-analysis because of missing values in the primary studies. Including all estimates would increase the statistical power of the analysis.

This chapter describes methods for data conversion, imputation and transformation based on the information published from the primary studies to enable the pooling of data from the different studies.

CHAPTER 6

St. John's wort for depression: a meta-analysis

6.1 Background and Aims

Unlike other herbal products, the efficacy of St. John's wort has been widely studied in randomised controlled trials. Four meta-analyses of these trials concluded that St. John's wort was more effective than placebo and may be as effective as other standard antidepressants in the short-term treatment of mild to moderately severe depression (Linde et al., 1996; Linde and Mulrow, 2003; Kim et al., 1999; Whiskey et al., 2001; William et al., 2000). However, almost all the meta-analysts emphasised the need for more studies particularly for trials comparing St. John's wort with standard antidepressants, as data are still insufficient to establish whether they are of similar effectiveness. Additionally, the authors recommended caution in interpretations of the results because of the heterogeneity of the RCTs in terms of trial quality, participants and interventions.

Since publications of these meta-analyses, six more RCTs investigating the effectiveness of St. John's wort for depression have been published (Kalb et al., 2001; Shelton et al., 2001; Behnke et al., 2002; Hypericum Depression

Trial Study Group, 2002; Lecrubier et al., 2002; van Gurp 2002). Three of these trials are large with a sample size of more than 200 participants.

Our aim is to update the meta-analyses.

6.2 Methods

6.2.1 Identification, selection of articles and data extraction

Randomised double blind trials were identified and selected as described in Chapter 4 (Section 4.2).

Two types of data were extracted (if available) from each study. The first type of data was dichotomous data, respond versus not respond. The second was continuous data, in the form of HAMD scores.

For continuous data, the methods for data conversion, imputation and transformation based on the information published from the primary studies to enable the pooling of data for the different studies were as described in Chapter 5 (Section 5.2).

6.2.2 Data synthesis

Data was analysed using Statsdirect (Version 2.2.3). We pooled results separately for four comparisons.

a. St. John's wort versus placebo (response rate)
b. St. John's wort versus placebo (change-from-baseline)
c. St. John's wort versus antidepressant (response rate)
d. St. John's wort versus antidepressants (change-from-baseline)

For trial with three treatment arms (Laakmann et al., 1998), we grouped together the two St. John's wort arms (0.5 and 5% hyperforin group) and compared them collectively with the placebo. Where trials had two control arms [Hypericum Depression Trial Study Group, (2002) and Philipp et al., (1999)] and the usual arm for treatment (St. John's wort), we considered each control arm as separate trial versus St. John's wort. i.e. trials of St' John's wort versus placebo and St. John's wort versus standard antidepressant.

For each comparison, the pooled estimates with 95% CI was calculated using both fixed and random effects models as appropriate. Pooled estimates were preceded by heterogeneity testing with an alpha level of 0.10.

6.2.3 Assessment of trial's quality

The number of methods for quality assessment of trials has been estimated to be between 50 and 60 (Verhagen et al., 2001). However, there is no agreement on how the quality of primary studies should be assessed and how the assessment should be incorporated into reviews (Moher et al., 1995). One approach in quality assessment has been to focus on important components such as randomisation and blinding (Moher et al., 1996; Schulz et al., 1995). Since our inclusion criteria required that only randomised, controlled and double blind studies were selected, we assumed that all these trials have results with internal validity. Jüni et al., (2001) described internal validity as a causal relationship between the experimental treatment (independent variable) and the observed effect (dependent variable). With regards to concealment of randomisation, we assumed that all trials have taken adequate measure to conceal allocation, even though the authors have not all described this aspect in detail.

6.2.4 Methods for pooling data

The primary outcome measure was whether the patient responded. For most trials, responders are based on the criteria of ≥ 50% reduction in HAMD score from baseline or a total score of ≤ 10 at the end of treatment. The secondary outcome measure was change-from-baseline (HAMD).

Dichotomous data

The dichotomous data for meta-analysis has been summarised using odds ratio (OR), the relative risk (RR), and the risk difference (RD) (Appendix 7). Both the OR and RR are relative measures and are used to combine studies, whereas the RD is an absolute measure and is useful when applied to a particular healthcare situation (Egger et al., 1997b). The OR has more attractive statistical properties (Fleiss, 1993). The OR and RR are similar if the outcome is relatively rare. The inverse of the RD provides the number needed to treat (NNT), which is helpful for clinicians, enabling them to translate the results to use in routine clinical practice (Osiri et al., 2003). However, it is generally accepted that RD is not an appropriate measure for use in meta-analysis (Altman, 2000).

We chose RR to express the effect of the intervention for our analysis because RR is relatively easier to interpret compared to OR. In particular, the OR has been criticised for not being well understood by physicians and patients (Sinclair and Bracken, 1994). Furthermore the appropriateness of OR to summarise data to express the statistical results of meta-analyses is still being debated (Khan et al., 1996).

Continuous data

Continuous data may be summarised using the method of weighed mean difference (WMD) and standardised mean difference (SMD), which is also called effect size (Appendix 7).

We used the WMD (Der Simonian and Laird, 1986) because all included trials measured outcomes on the same scale (HAMD). The weight given to each study was determined by the accuracy of its effect estimate. In the statistical software used by us (Statsdirect Version 2.2.3), this is equal to the inverse of the variance.

We pooled the result of continuous data using the change-from-baseline score. Mean change-from-baseline was defined as the difference between the baseline and endpoint score. This value reflects the degree of change in symptoms levels from baseline, measured using HAMD rating scales. Where data on mean change-from-baseline were not available, this value was calculated (as described in Chapter 5 (section 5.2) from the endpoint HAMD values reported in the trial.

6.2.5 Test for statistical heterogeneity

Statistical heterogeneity of treatment effect was assessed using the Q statistic, which approximates the χ^2 distributions with n-1 degrees of freedom (DerSimonian and Laird, 1986). This test examined whether the observed variability in study results (among the trials) was greater than expected by chance. A large value of Q (low probability of occurrence) indicates that there is significant heterogeneity between studies. Berlin et al. (1989) noted that this test has low power and not very good at detecting heterogeneity when there are few studies. For this reason, we set the significance level of this statistic at 0.10 rather than the usual 0.05 because there were few studies involved in our review.

Fixed effects model assumes that an intervention has a single true effect, whereas random effects model assumes that an effect may vary across studies. We presented the results from both fixed and random effects model in the tables. The random effects model was expected to give a more conservative estimate (wider confidence intervals).

On the basis of Q statistic for heterogeneity, if the heterogeneity test was statistically significant, then a random effects model was used for making inferences.

6.2.6 Investigating bias

We examined the presence or absence of publication bias graphically using funnel plots. Funnel plots are simple scatter plots of the treatment effects estimated from individual studies against some measure such as sample size or standard error. Both the sample size and the number of events (participants responding to treatments) determined the statistical power of a study (Sterne and Egger, 2001). As such it was reasonable to choose standard error (SE) as the vertical axis for our plot. Using the SE rather than the inverse of SE emphasised the differences between studies of smaller size, for which biases are more likely to be found (Sterne and Egger, 2001).

A funnel plot is based on the fact that the precision in estimating treatment effect increases as the sample size increases. Therefore, effect estimates from small studies will scatter more widely at the bottom of the plot, with less spread among larger studies. Asymmetrical plots suggest that biases are present.

6.3 Results

6.3.1 Studies included

A total of 30 relevant studies were identified. The characteristics of these trials in terms of the demographics of patients and trial settings; diagnostic classifications and assessment durations; types of St. John's wort preparations and daily dose were as presented in Chapter 4 (Table 4.2; Table 4.3; Table 4.4 respectively).

6.3.2 List of comparisons

a. St. John's wort versus placebo (response rate)

Eighteen trials involving 1686 patients contributed data to the analyses comparing the efficacy of St. John's wort with placebo (Table 6.1). Dosages of St. John's extract varied from 300mg to 1800 mg daily. One trial (Harrer et al., 1991) was omitted from the analysis because data on response rate was not available.

Table 6.1 St. John's wort versus placebo (response rate)

Study	St. John's wort (n/N)	Placebo n/N	Weight (%)	Relative Risk (95% CI)
Halama (1991)	10/25	0/25	0.26	21.0 (1.30 - 340.00)
Hängsen et al. (1996)	35/53	12/54	5.94	2.97 (1.80 - 5.15)
Hübner et al. (1994)	14/20	9/20	4.50	1.56 (0.91 - 2.85)
Hypericum Depression Trial Study (2002)	16/113	13/116	6.41	1.26 (0.65 - 2.48)
Kalb (2001)	23/37	15/35	7.71	1.45 (0.93 - 2.34)
Laakmann et al. (1998)	48/98	16/49	10.67	1.34 (0.87 – 2.17)
Lecrubier (2002)	98/186	80/189	39.68	1.24 (1.00 - 1.55)
Lehrl et al. (1993)	4/25	2/25	1.00	2.00 (0.47 - 8.82)
Philipp et al. (1999)	67/100	22/46	15.07	1.40 (1.04 - 2.01)
Quandt et al. (1993)	29/44	4/33	2.29	5.43 (2.33 - 13.95)
Reh et al. (1992)	20/25	11/25	5.50	1.82 (1.16-3.08)
Schlich et al. (1987)	15/25	3/24	1.53	4.80 (1.79-14.32)
Schmidt & Sommer (1993)	20/32	6/33	2.95	3.44 (1.69-7.54)
Schmidt et al. (1989)	10/20	4/20	2.00	2.50 (1.01-6.69)
Schrader et al. (1998)	45/81	12/81	6.00	3.75 (2.20-6.60)
Shelton et al. (2001)	26/98	19/102	9.31	1.42 (0.85-2.40)
Sommer & Harrer (1994)	28/50	13/55	6.19	2.37 (1.42-4.09)
Witte et al. (1995)	34/48	25/49	12.37	1.39 (1.01-1.97)
Fixed Effects				1.79 (1.59-2.00)
Random Effects				1.90 (1.54-2.35)

Test for homogeneity Q = 45.33 with (df = 17), P = 0.0002

Responders are based on the criteria of ≥ 50% reduction in HAMD score from baseline or a total score of ≤ 10 at the end of treatment.

A visual examination of the Forrest plot of the meta-analysis (Figure 6.1) shows a clear variation in treatment effects between the RCTs. Each line on the plot shows the point estimate effect (■) and the 95% confidence interval. Another graphical exploration by the use of L'Abbé plot (Figure 6.2) shows that all of the trials were in the upper left half of the graph, demonstrating effectiveness of St. John's wort. However, it can be seen that the responder rates vary greatly in both the treatment (14.2% to 80.0%) and the placebo (0.0% to 51.0%) groups indicating heterogeneity of effects among the different trials.

The homogeneity test further suggests significant heterogeneity (Q = 45.33 with 17 df, p = 0.0002). As such the fixed effects model was not suitable. Using the random effects model, St. John's wort was shown to be significantly more effective than placebo with a pooled RR of 1.90 (1.54- 2.35) (Table 6.1 and Figure 6.1). Patients on St. John's wort were almost twice as likely as placebo to respond.

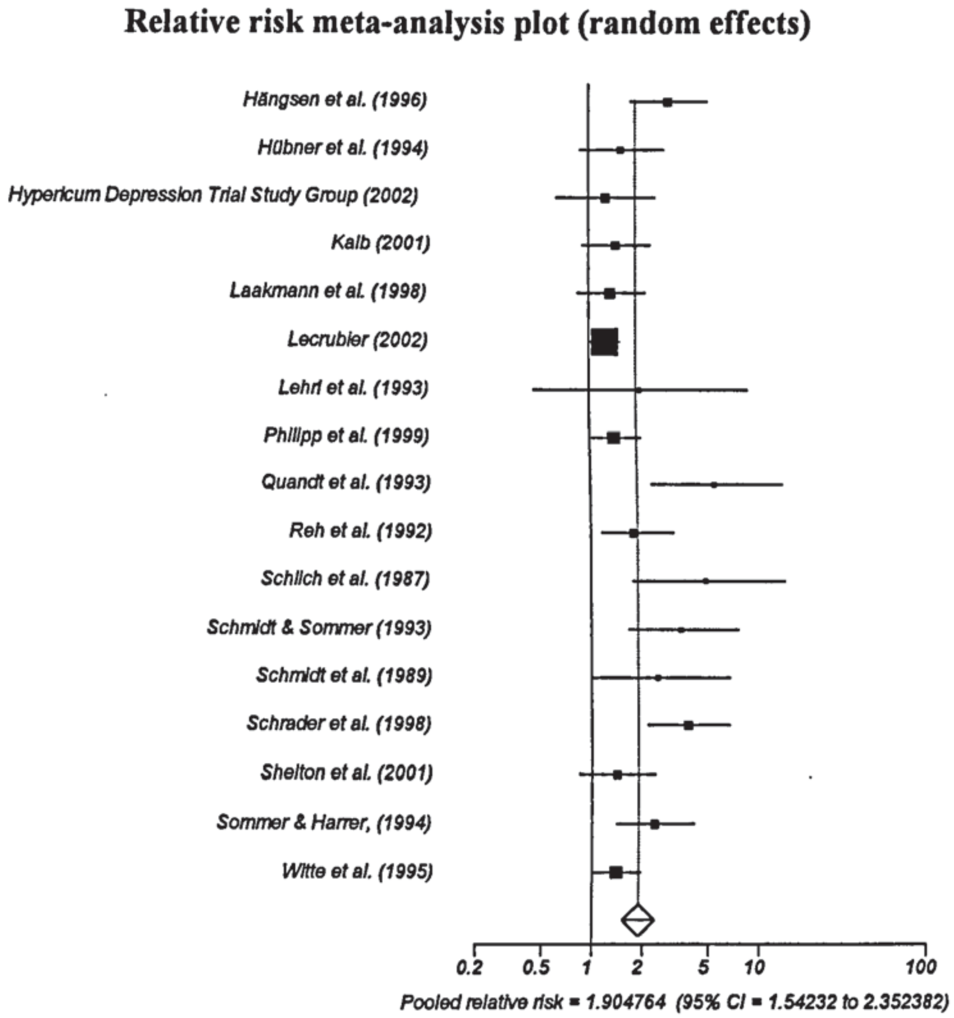


Figure 6.1 Plot of pooled efficacy of St. John's wort compared with placebo (response rate). Results falling to the right of the line of no effect (one) indicate that St. John's wort was more effective.

The squares give the point estimates and the horizontal line across each point gives the 95% confidence interval. The size of the square represents the weight assigned to the study concerned.

The pooled estimate is shown as a diamond shape. The vertical dotted line is the pooled estimate of effect.

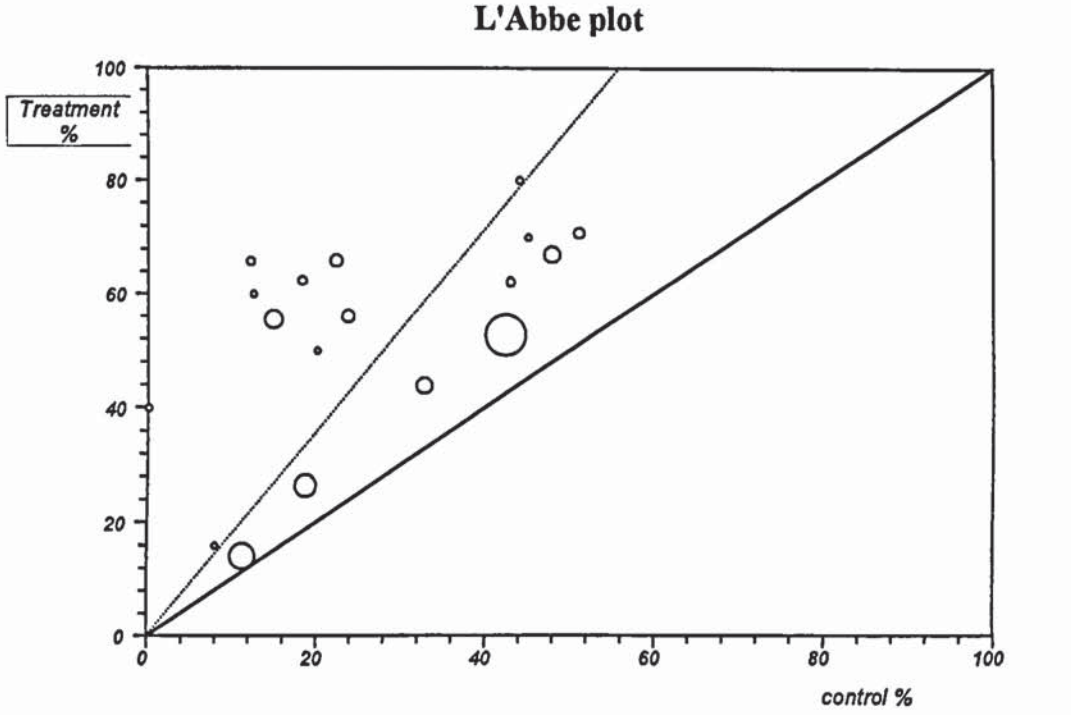
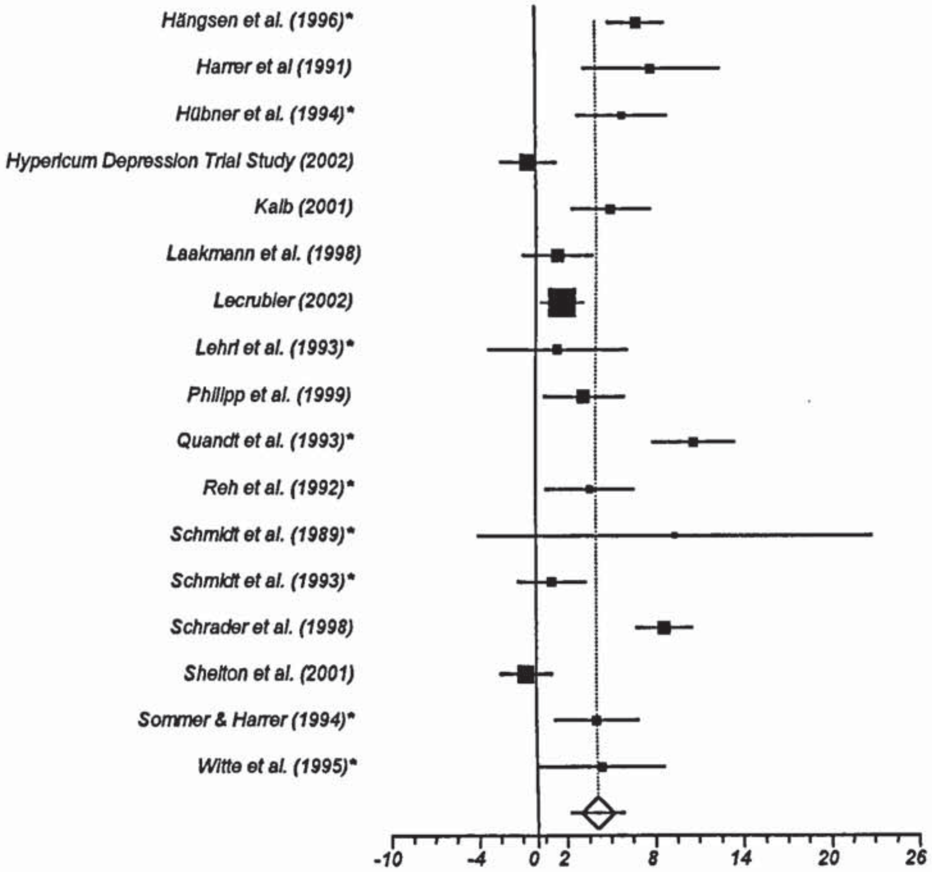


Figure 6.2 L'Abbe plots of placebo-controlled trials of St. John's wort in depression. Solid diagonal line represents the line of equality of event (responder) rates in the two arms within trials. The dotted line (the overall RR line) represents a summary RR of 1.90. The size of circles gives the relative weight contributed by the trial concerned.

b. St. John's wort versus placebo (change-from-baseline)

In the analysis of St. John's wort with placebo using change-from-baseline score, there were 17 comparisons (Figure 6.3). Two trials (Halama, 1991; Schlich et al., 1987) were omitted from the analysis because data on change-from-baseline score was not available. The homogeneity test suggested significant heterogeneity (Q = 130.77 with 16 df, p < 0.0001). As such the fixed effects model was not suitable. Using the random effects model, the pooled WMD was 4.09 (95% CI 2.33 to 5.84) in favour of St. John's wort, suggesting a reduction of around 4 points from baseline in those receiving St. John's wort (Figure 6.3).

Effect size meta-analysis plot (random effects)



Pooled WMD = 4.085184 (95% CI = 2.332989 to 5.83738)

* Data were derived from endpoint values.

Figure 6.3 Plot of pooled efficacy of St. John's wort compared with placebo (change-from-baseline). Results falling to the right of the line of no effect (zero) indicate greater efficacy for St. John's wort.

For explanations of the symbols see legends under Figure 6.1 page 105.

c. St. John's wort versus antidepressants (Response rate)

Twelve trials involving 1920 patients contributed data to the analyses comparing the efficacy of St. John's wort with antidepressants (Table 6.2). The various antidepressants used as comparators were amitriptiline (Bergman et al., 1993; Wheatley, 1997); fluoxetine (Behnke, 2002; Harrer et al., 1999; Schrader, 2000), imipramine (Phillip et al., 1999; Vorbach et al., 1994; Vorbach et al., 1997; Woelk, 2000), maprotiline (Harrer et al., 1994), and sertraline (Brenner et al., 2000; Hypericum Depression Trial Study Group, 2002). One trial (van Gurp et al., 2002) was omitted from the analysis because response rate data was not available for this trial.

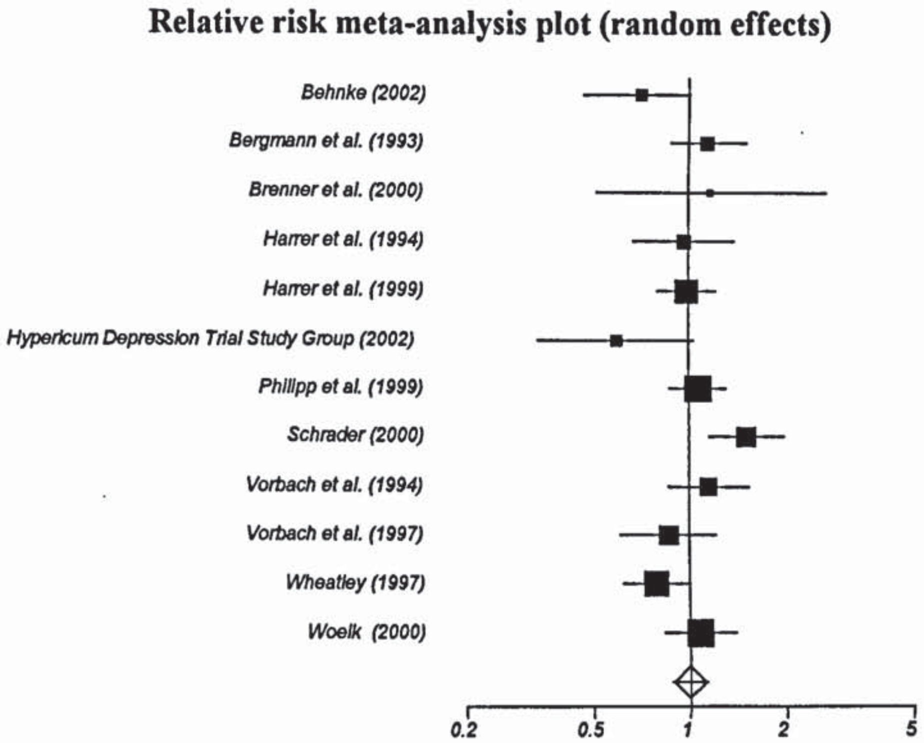
Table 6.2 St. John's wort versus antidepressants (response rate)

Study	St. John's wort (n/N)	Anti-depressants (n/N)	Weight (%)	Relative Risk (95% CI)
Behnke 2002	16/29	25/32	11.89	0.71 (0.47-1.01)
Bergmann et al. (1993)	32/40	28/40	14.00	1.14 (0.88-1.51)
Brenner et al. (2000)	7/15	6/15	3.00	1.17 (0.52-2.69)
Harrer et al. (1994)	27/51	28/51	14.00	0.96 (0.67-1.39)
Harrer et al. (1999)	50/70	57/79	26.78	0.99 (0.80-1.21)
Hypericum Depression Trial Study Group (2002)	16/113	26/109	13.23	0.59 (0.34 – 1.03)
Philipp et al. (1999)	67/100	66/105	32.20	1.07 (0.87 – 1.31)
Schrader (2000)	75/125	45/113	23.63	1.51 (1.16-1.99)
Vorbach et al. (1994)	42/67	37/68	18.36	1.15 (0.87-1.54)
Vorbach et al. (1997)	38/107	42/102	21.50	0.86 (0.61-1.22)
Wheatley (1997)	50/87	57/78	30.05	0.79 (0.62-0.98)
Woelk (2000)	68/157	67/167	32.47	1.08 (0.83-1.40)
Fixed effects				1.01 (0.93 –1.10)
Random effects				1.00 (0.88 –1.13)

Test for homogeneity Q = 23.44 with (df = 11), p = 0.015

Responders are based on the criteria of $\geq$ 50% reduction in HAMD score from baseline or a total score of $\leq$ 10 at the end of treatment.

A visual examination of the Forrest plot of the meta-analysis comparing St. John's wort with antidepressants (Figure 6.4) shows variation in treatment effects between the RCTs. Another graphical exploration by the use of L'Abbé plot (Figure 6.5) shows the line of equality, almost overlaps the overall RR line (dotted line), suggesting that both St. John's wort and antidepressants were of similar effectiveness. However, it can be seen that the responder rates vary greatly in both the St. John's wort (14.2 to 80.0%) and the antidepressants (23.9 to 73.0%) groups suggesting heterogeneity of effects among the different trials. The homogeneity test suggested significant heterogeneity (Q = 23.44 with 11 df, p = 0.015). Using the random effects model, a pooled RR of 1.00 (0.88 – 1.13) was obtained (Table 6.2 and Figure 6.4), suggesting that St. John's wort was as likely as conventional antidepressants to improve depression.



Pooled relative risk = 0.999264 (95% CI = 0.882043 to 1.132062)

Figure 6.4 Plot of pooled efficacy of St. John's wort compared with antidepressants (response rate). Results falling to the right of the line of no effect (one) indicate an advantage for St. John's wort.

The squares give the point estimates and the horizontal line across each point gives the 95% confidence interval. The size of the square represents the weight assigned to the study concerned. The pooled estimate is shown as a diamond shape.

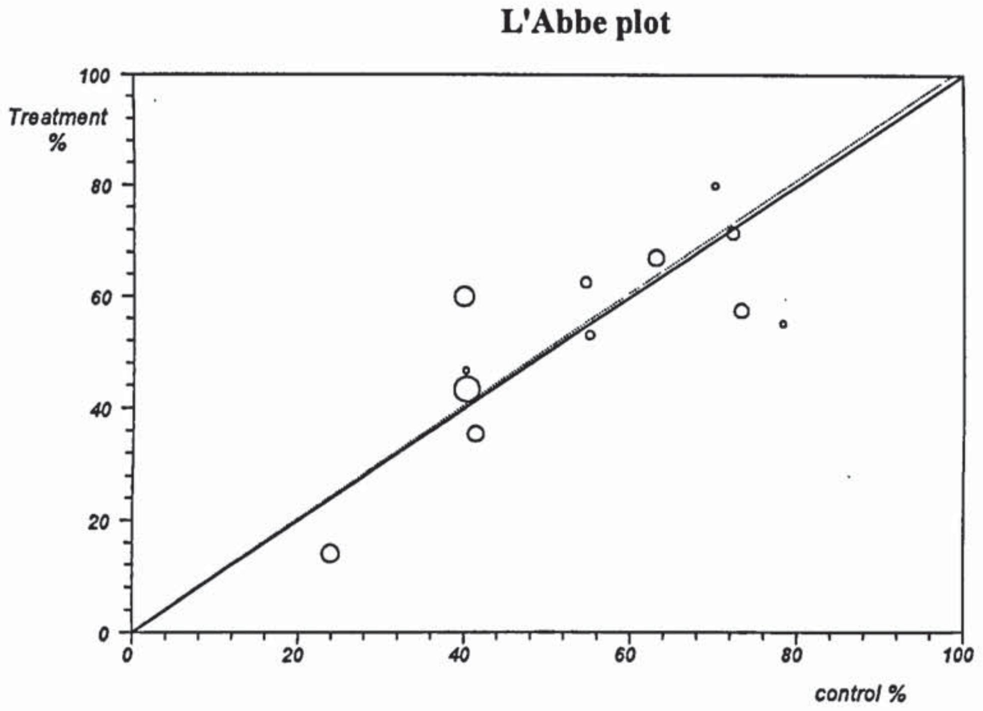


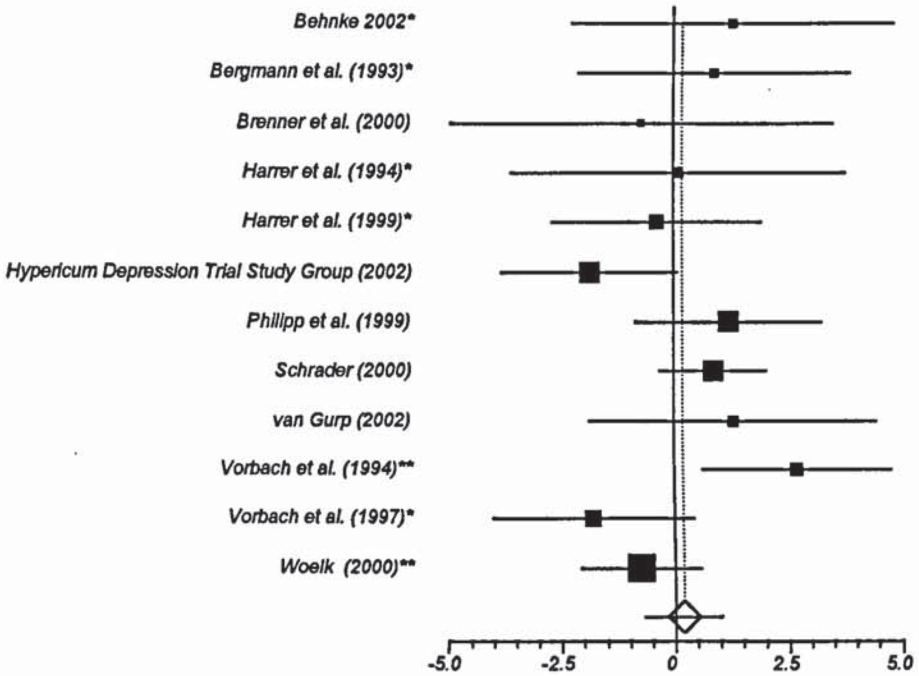
Figure 6.5 L' Abbe plot of antidepressant-controlled trials of St. John's wort in depression. Solid diagonal line represents the line of equality of event (responder) rates in the two arms within trials. The size of circles gives the relative weight contributed by the trial concerned.

d. St. John's wort versus antidepressants (change-from-baseline)

In the analysis of St. John's wort with antidepressants using change-from-baseline data, there were 12 comparisons (Figure 6.6). As currently there is no reliable methods available to conduct a meta-analysis using medians, one trial (Wheatley, 1997) was omitted from the analysis.

It can be seen from Figure 6.6 that none of the trials except that by Vorbach et al. (1994) demonstrated a statistically significant difference between the two treatment groups either individually or once pooled. The homogeneity test suggested significant heterogeneity (Q = 18.56 with 11 df, p =0.07). Using the random effects model, the WMD for mean change was 0.18 (95% CI –0.66 to 1.02), suggesting little difference in reduction of symptoms (HAMD scores) from baseline observed in the St. John's wort and conventional antidepressant groups.

Effect size meta-analysis plot (random effects)



Pooled WMD = 0.178859 (95% CI = -0.658647 to 1.016365)

* Data were derived from the endpoint values
** Required data was imputed.

Figure 6.6 Plot of efficacy of St. John's wort compared with conventional antidepressant (change-from-baseline). Results falling to the right of the line of no effect (zero) indicate an advantage for St. John's wort.

The squares give the point estimates and the horizontal line across each point gives the 95% confidence interval. The size of the square represents the weight assigned to the study concerned. The pooled estimate is shown as a diamond shape. The vertical dotted line is the pooled estimate of effect.

6.3.3 Publication bias

a. St. John's wort versus placebo

Figure 6.7 shows an asymmetrical funnel plot (the plot is weighted to the right), suggesting the likelihood of bias towards positive trials reporting results in favour of St. John's wort.

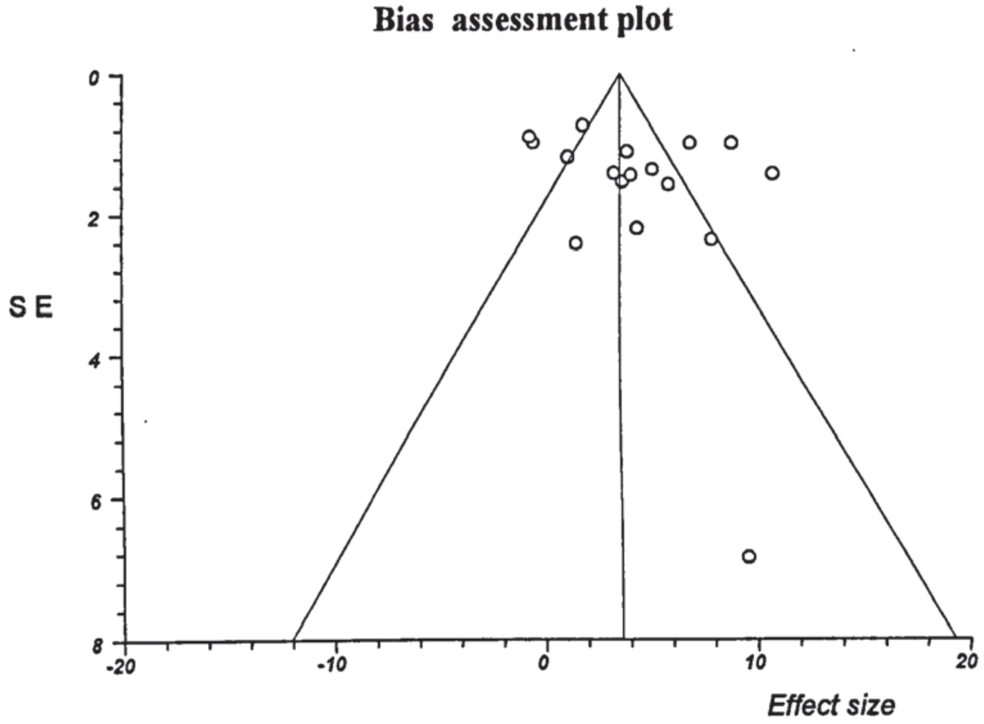


Figure 6.7 Funnel plot of trials comparing St. John's wort with placebo

b. St. John's wort versus antidepressants

As can be seen from Fig 6.8 the plot is asymmetrical, suggesting the likelihood of bias.

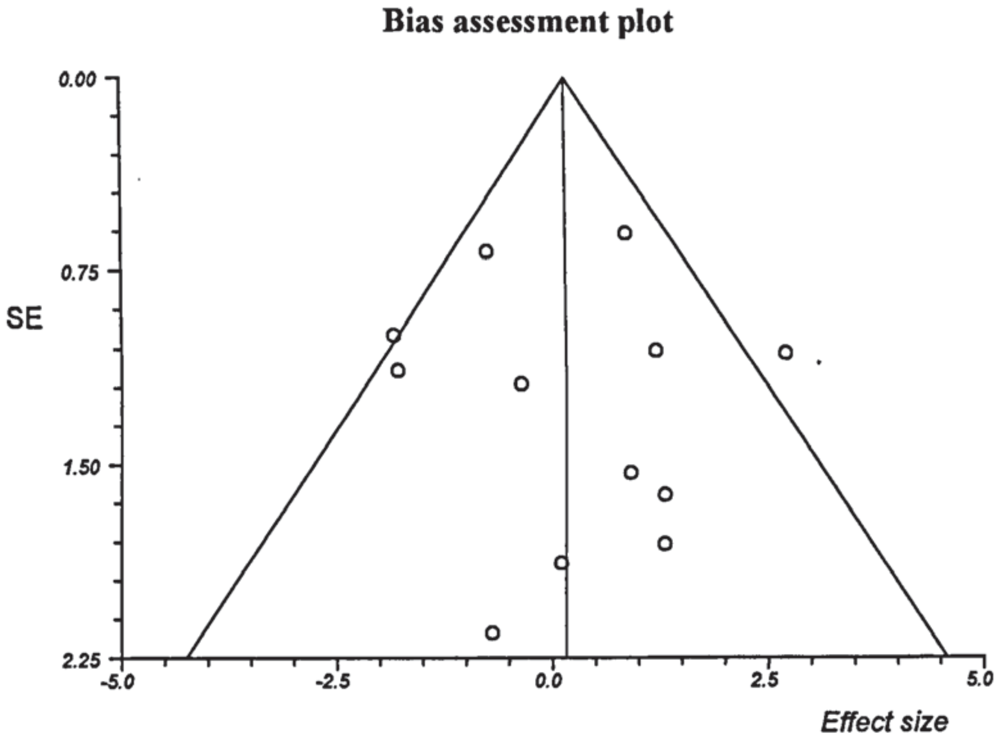


Fig. 6.8 Funnel plot of trials comparing St. John's wort with antidepressants

6.4 Discussion

St. John's wort versus placebo

The result of this meta-analysis is generally in agreement with the results of the meta-analysis by Linde and Mulrow (2003). However, we found that St. John's wort was only 1.9 times as likely as placebo to improve depression, rather than almost 2.5 times suggested by Linde and Mulrow (2003). Our value was lower because of the contributions of data from three recent large and more methodologically rigorous trials (Shelton et al., 2001; Hypericum Depression Trial Study Group, 2002; Lecrubier et al., 2002), of which two (Shelton et al., 2001; Hypericum Depression Trial Study Group, 2002) did not support the claim of antidepressive efficacy for St. John's wort.

Pooling HAMD data suggests that on average St. John's wort decreases the score by 4.0 HAMD points over placebo. Montgomery (1994) has suggested that such a difference between treatment and control group may be important. However, there is no consensus on what represents a clinically significant improvement on the HAMD scale.

This meta-analysis using response rate provides some evidence that St. John's wort is more effective than placebo. However, the majority of the studies were sponsored by manufacturers and the possibility that sponsorship may bias reporting must be borne in mind (Stewart and Parmar, 1996; Freemantle et al., 2000).

St. John's wort versus antidepressants

Our result confirms previous findings (Linde et al., 1996; Linde and Mulrow, 2003; Kim et al., 1999; Whiskey et al., 2001; William et al., 2000) that St. John's wort has similar efficacy to conventional antidepressants. However, with this type of non-inferiority trials (new treatment versus active controls), equivalence has not been demonstrated. Though unlikely, both treatments may be equivalent and ineffective.

Heterogeneity of Effects

Given that the RCTs of St. John's wort for depression differed considerably in terms of patients' characteristics, types of interventions and length of trials (Chapter 4), it is not surprising to find the evidence of heterogeneity from the graphical explorations and test of heterogeneity. As the evidence for heterogeneity of effects is substantial, this introduces doubt about the interpretations of the pooled estimate of effect, unless the sources of heterogeneity can be explained. These will be explored in the next Chapter (Chapter 7).

Publication Bias

Funnel plots are useful for detecting publication bias in meta-analysis (Egger et al., 1997a). Asymmetrical plots as seen in both our meta-analyses are suggestive of publication biases. However, since asymmetry is generally defined informally, through visual examination, the visual interpretation of funnel plots may vary between observers (Villar, 1997).

Publication bias could be an issue in the interpretation of pooled estimate of treatment effect because there are several studies of small size sponsored by the manufacturers, suggesting that positive results of St. John's wort treatment were systematically reported and that the treatment effect may be overestimated. Even though publication bias has been generally associated with funnel plot asymmetry (Song et al., 2002), it is acknowledged that funnel plot asymmetry could also be caused by other factors such as poor methodological quality of studies, choice of effect measure or a play of chance (Egger et al., 1997a). Randomisation and concealment of allocation have been shown to influence outcomes in RCT. Although all the trials in this review reported their treatment assignment procedure as being randomised, not all trials described the randomisation procedure in detail.

Since our meta-analyses are based on a limited number of small trials, the result of the funnel plot analysis should be treated with considerable caution.

6.5 Conclusion

The present analysis, which included six additional recent RCTs provide some evidence to suggest that St. John's wort is useful in patients with depression. However, because of heterogeneity and possible publication bias in the RCTs of St.John's wort overemphasis of any single pooled estimate of effect may be misleading. Understanding the reasons for the observed heterogeneity is necessary before any definitive conclusion from these data can be made.

CHAPTER 7

Exploring sources of heterogeneity in placebo-controlled trials of St. John's wort for depression

7.1 Background and aim

One of the major criticisms directed towards meta-analysis of published studies is the pooling of heterogeneous studies, a practice which makes interpretation of any resultant pooled estimate of effect difficult. This problem is sometimes referred to as the mixing of apples and oranges. However, given that clinical studies, even those aiming to be replicates of each other, are never fully identical to each other, the basic problem is not per se, the pooling of results from studies with some differences in designs, but a loss of sight of factors which contribute to observed differences in effect. So, without a clear view of potential sources of heterogeneity, the applicability of the results to given practical scenarios is difficult to gauge. Investigating possible sources of variation can lead to important insights about treatment effects as commented by Colditz et al. (1995):

"In a meta-analysis, documenting heterogeneity of effect (by identifying sources of variability in results across studies) can be as important as reporting averages. Heterogeneity may point to situations in which an intervention works and those in which it does not. Finding systematic variation in results and identifying factors that may account for such variation, in this way, aids in the interpretation of existing data and the planning and execution of future studies."

Placebo controlled trials of St. John's wort for depression differed considerably such as in diagnostic depression criteria, types of St. John's wort preparation and length of trials (Chapter 4, section 4.3). As such, it would not be surprising to find that the estimates of treatment effect of these trials were different from one another. This kind of variations in quantitative results between the trials is termed statistical heterogeneity (Thompson, 1994). The graphical explorations and test of homogeneity in the meta-analysis of St. John's wort versus placebo described in Chapter 6 show substantial evidence for statistical heterogeneity. As such simply pooling the results of these trials into one overall summary estimate may be misleading.

We chose randomised controlled trials of St. John's Wort in the treatment of depression because the herb is used by both practitioners of alternative medicine, and by conventionally trained physicians. There is therefore a potential clash in the perspectives of these two groups of carers, in relation to both management of patients and perception of the determinants of success in a clinical trial, both of which may contribute to heterogeneity in treatment effects. Moreover, there is as yet no consensus as to whether St. John's Wort is effective in depression.

In this study we aimed to identify factors, which could potentially contribute to heterogeneity of effects in placebo-controlled trials of St. John's wort. The examination of heterogeneity should increase the relevance of the conclusion drawn from our meta-analysis in comparing St. John's wort versus placebo.

7.2 Methods

7.2.1 Identification, selection of trials and data extraction

We explored the heterogeneity in the meta-analysis of St. John's wort versus placebo. Trials identified and included in the meta-analysis in comparing St. John's wort versus placebo (Chapter 6, Section 6.3.2a) were used for the assessment.

The following data were extracted: study identifications, diagnostic classification of depression, and length of study, dichotomous outcomes and continuous outcomes. I extracted the data and my supervisor cross-validated them.

7.2.2 Subgroup analysis

The method used frequently to explore heterogeneity is subgroup analysis. Studies are categorised according to the characteristics of the trial and a summary estimate of effect is obtained for trials in each of the categories. We hypothesised two main sources for heterogeneity: (a) treatment effect differs according to the different diagnostic classification of depression (b) treatment effect differs according to different length of follow-up.

a. Diagnostic classification of depression

Two groups of trials were compared: those using the earlier classification system (ICD-9) and those using the more recent classification criteria (DSM – IIIR, DSM-IV, and ICD-10).

Diagnostic classification was examined because diagnostically, depressive disorders cover a wide group, which includes major depression as well as depressive states not satisfying the criteria for the full syndrome. For trials, which used ICD-9 classification, the patients would have closely matched the DSM-IV categories of adjustment disorder with depressed mood (code: 309.0) or acute stress disorder (code: 308.3) rather than major depression (code 296.XX). Because such patients are depressed in response to a crisis or stressor, their symptoms remit once the stressor is removed (Casey, 2001). We therefore, anticipated that trials with this group of patients would show greater treatment effects compared to trials, which studied patients with major depression.

b. Length of trials

Two categories of trials were compared: those trials lasting 4 weeks and shorter and those trials lasting longer than 4 weeks.

We anticipated that length of trials would likely affect the treatment effects because it has been shown that depressive effects of antidepressants might be, present early in treatment, late or consistent over time (Gelber and Golhirsch, 1987). There is evidence that many antidepressants begin to show effects very early in treatment, and a significant difference from placebo may be seen as early as one week (Montgomery, 1995).

Within each group, test of equality of average score were undertaken with the nonparametric Mann-Whitney U test. We chose nonparametric statistics because our sample was small and the normality assumption was unlikely to be valid. Statistical analyses were performed with Statistical Package for the Social Sciences (SPSS), version 10.0 for Windows.

7.2.3 Loglinear analysis

Log-linear techniques are useful for uncovering the potentially complex relationships among variables in a multiway cross-tabulation (Gilbert, 1981). We used loglinear analysis to identify the interactions and associations of the three variables: diagnostic classification, length of trials and treatment effects. The method of backward hierarchical elimination was used to build the model.

Based on the meta-analysis of trials of St. John's wort versus placebo using response data (section 6.3.2a) we categorised the treatment effect into two categories: treatment effect which was statistically significant, and treatment effect which was not statistically significant. The variable, length of trials was dichotomised into two discrete categories: 4 weeks and less, and more than 4 weeks. The two categories for diagnostic classifications were: the early classification (ICD-9) and the recent classifications (DSM-IIIR, DSM-IV, ICD-10).

The analysis was performed with Statistical Package for the Social Sciences (SPSS), version 10.0 for Windows.

7.3 Results

7.3.1 Subgroup analyses

Even when trials were grouped according to diagnostic classifications and duration of trials the heterogeneity was still substantial except for two subgroups (Figure 7.1 and 7.3). A meta-analysis of other subgroups showed observed effects were not homogeneous within the subgroups (Figure 7.2, Figure 7.4 to 7.8).

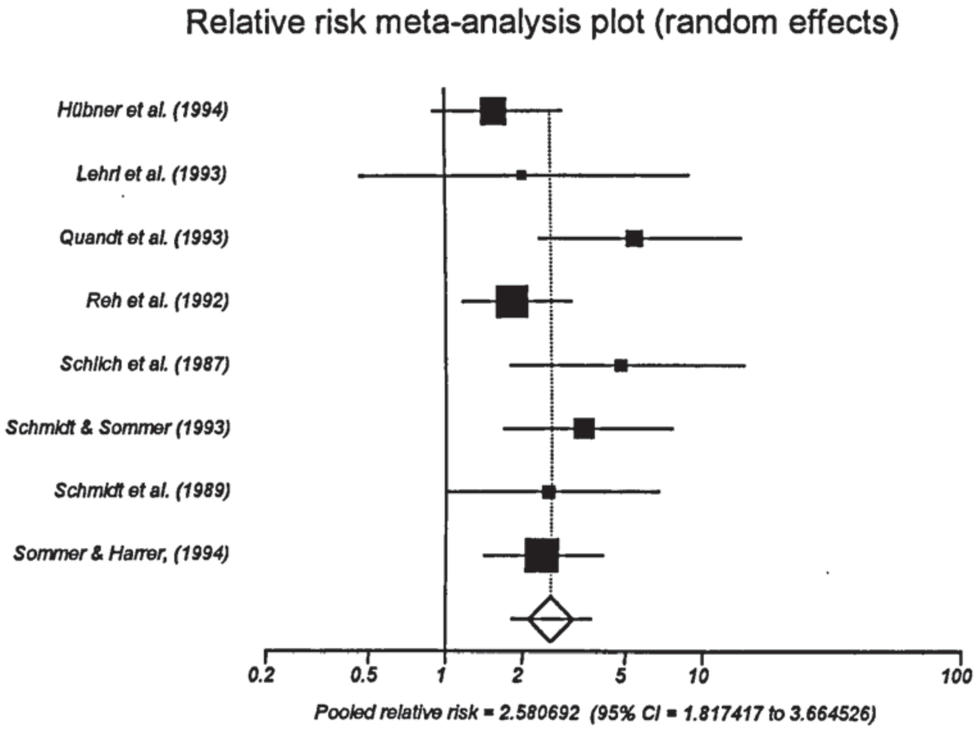


Figure 7.1 Subgroup analysis of trials with ICD-9 classification (response rate). The observed effects were statistically homogeneous with a significance probability. Test of heterogeneity (Q = 13.13 with df = 8, p = 0.1075).

The squares give the point estimates and the horizontal line across each point gives the 95% confidence interval. The size of the square represents the weight assigned to the study concerned. The pooled estimate is shown as a diamond shape. The vertical dotted line is the pooled estimate of effect.

Figure 7.2 shows that although the diagnostic instruments were apparently homogeneous (Appendix 3, 4, and 5), the observed effects were not (p for homogeneity = 0.007).

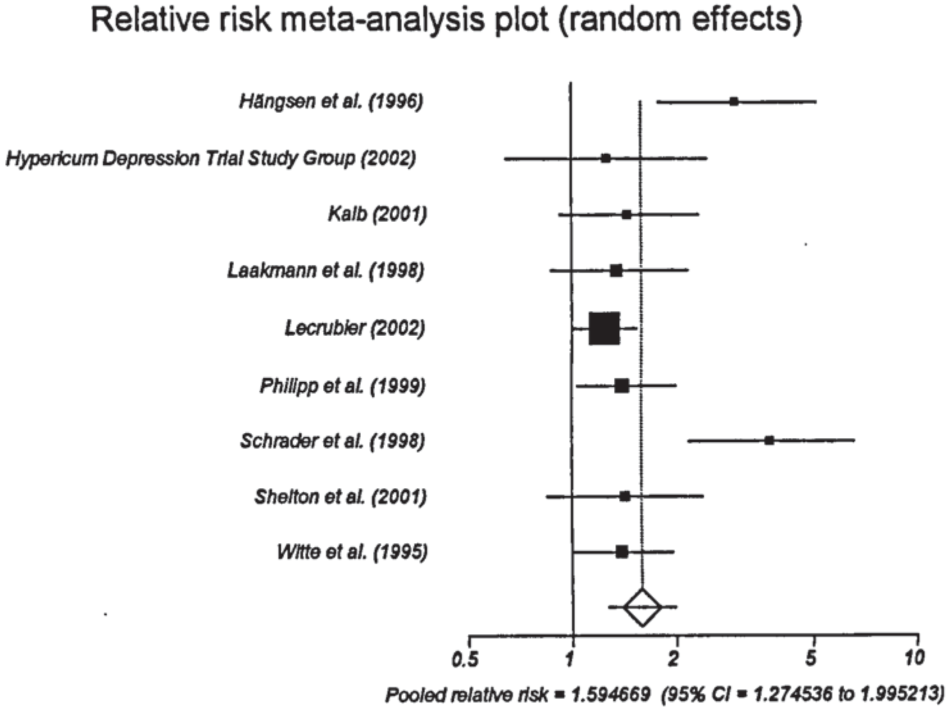


Figure 7.2 Subgroup analysis of trials with diagnostic classification (DSM-IIIR, DSM-IV, ICD-10) (response rate). The observed effects were statistically heterogeneous. Test of heterogeneity (Q = 20.92 with df = 8, p = 0.007).

The squares give the point estimates and the horizontal line across each point gives the 95% confidence interval. The size of the square represents the weight assigned to the study concerned. The pooled estimate is shown as a diamond shape. The vertical dotted line is the pooled estimate of effect.

Within the subgroup of trials which lasted 4 weeks and less, the observed effects were statistically homogeneous (Figure 7.3)

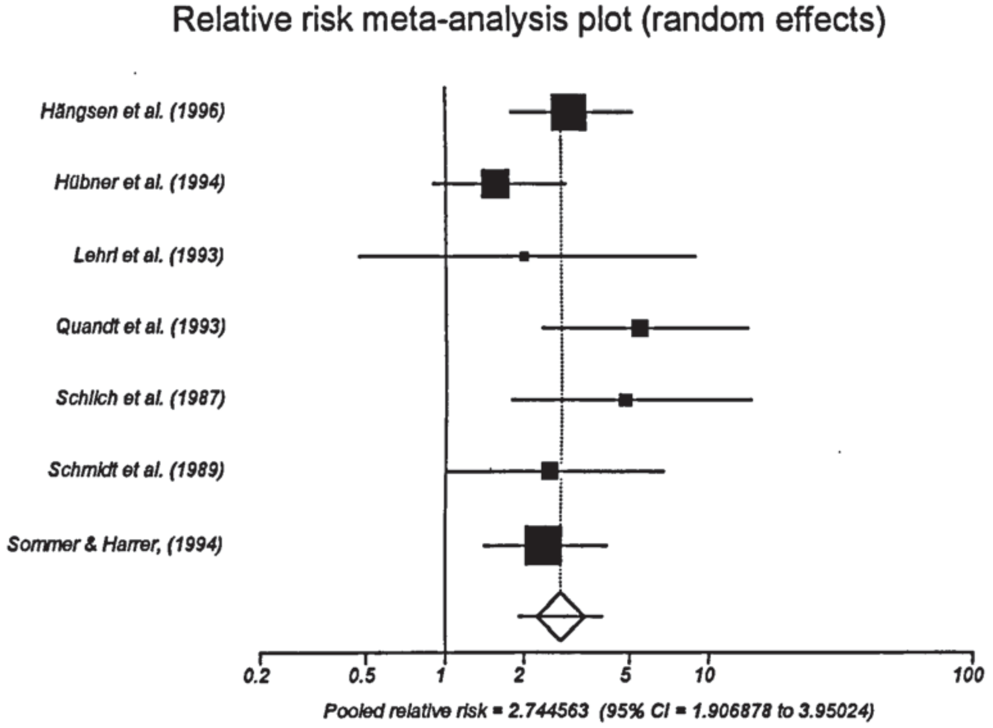


Figure 7.3 Subgroup analysis of trials lasting 4 weeks and less (response rate). The observed effects were statistically homogeneous. Test of heterogeneity (Q = 10.54 with df = 7, p = 0.1597).

The squares give the point estimates and the horizontal line across each point gives the 95% confidence interval. The size of the square represents the weight assigned to the study concerned. The pooled estimate is shown as a diamond shape. The vertical dotted line is the pooled estimate of effect.

However, similar results were not seen within the subgroup of trials lasting more than 4 weeks (Figure 7.4)

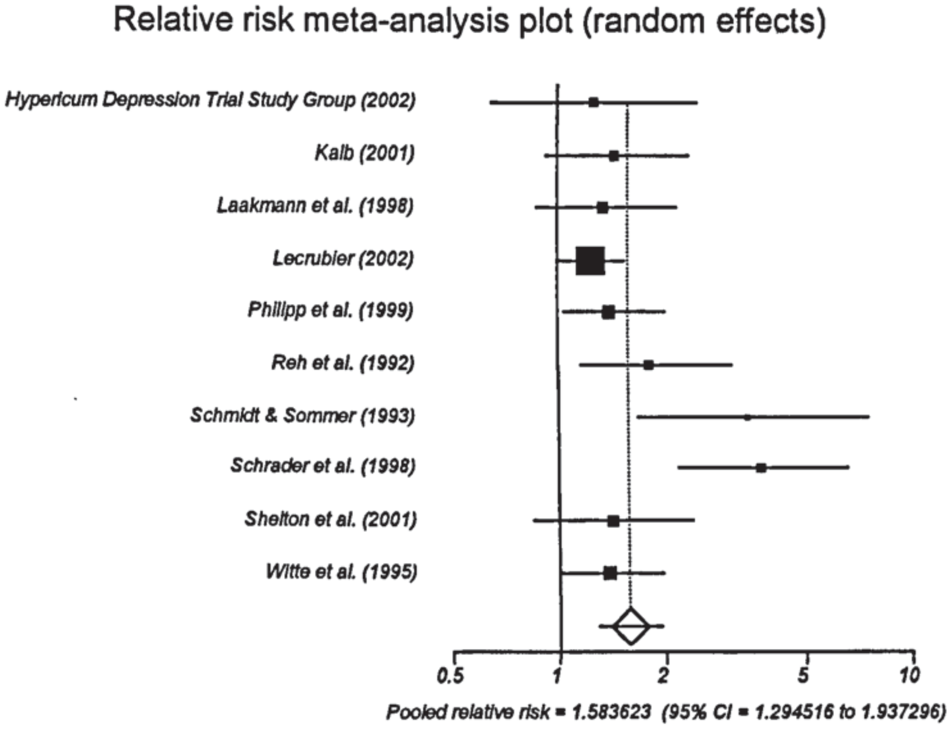
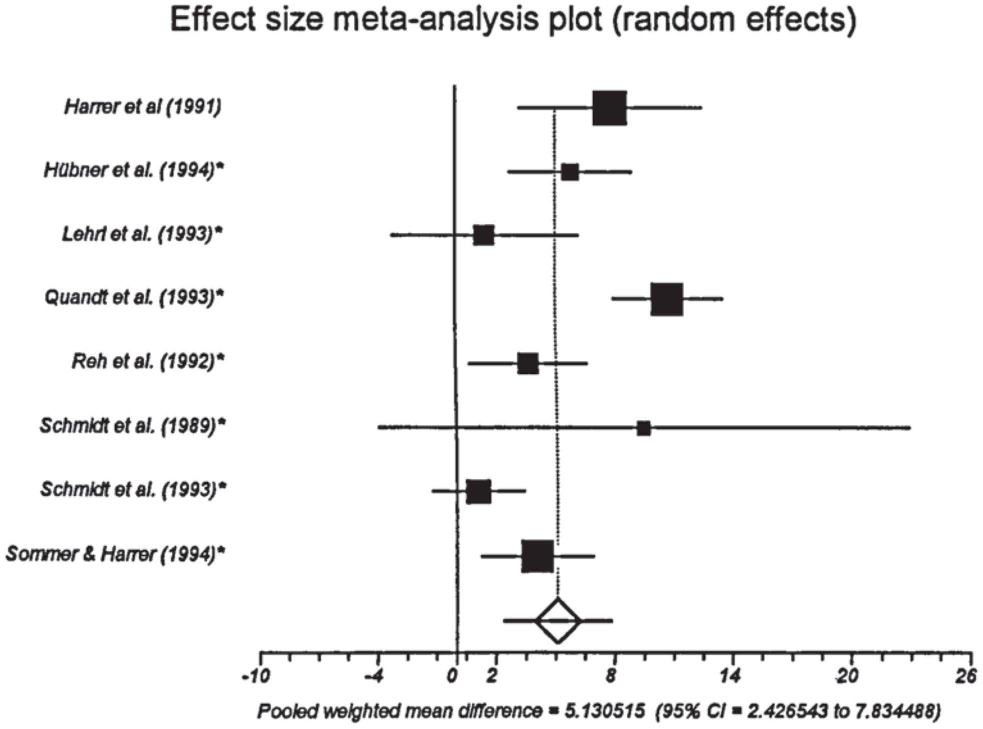


Figure 7.4 Subgroup analysis of trials lasting more than 4 weeks (response rate). The observed effects were statistically heterogeneous. Test of heterogeneity (Q = 19.55 with df = 9, p = 0.021).

The squares give the point estimates and the horizontal line across each point gives the 95% confidence interval. The size of the square represents the weight assigned to the study concerned. The pooled estimate is shown as a diamond shape. The vertical dotted line is the pooled estimate of effect.

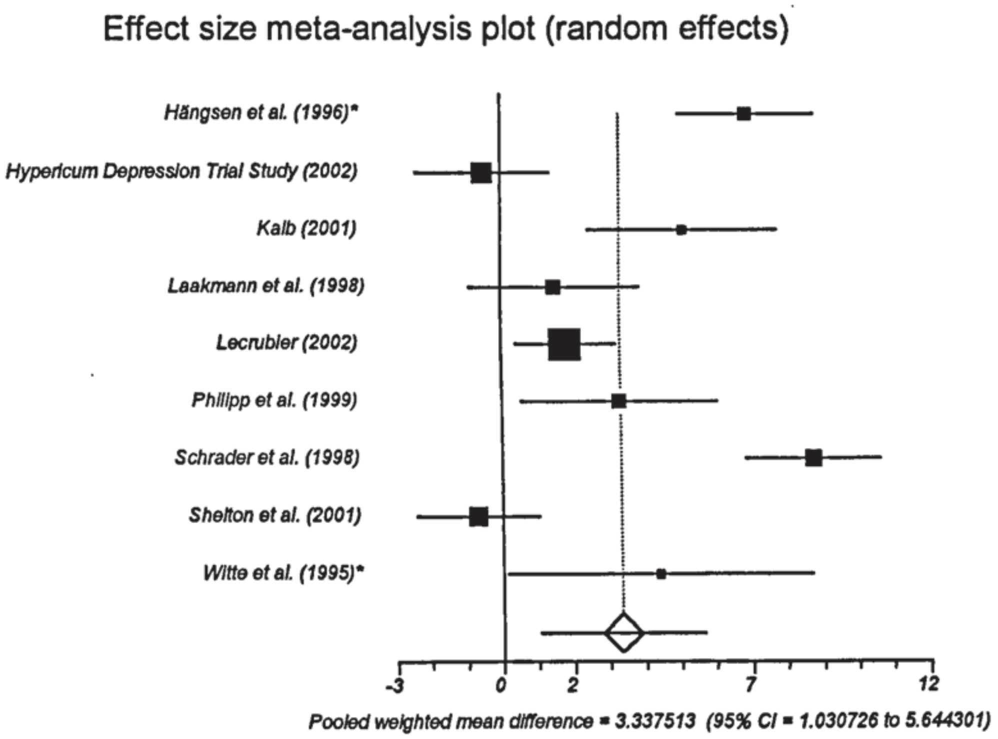
Heterogeneity of effects was observed within all the subgroups using change-from-baseline data (Figure 7.5 to Figure 7.8).



*Data derived from endpoint values.

Figure 7.5 Subgroup analysis of trials with ICD-9 diagnostic classification (change-from-baseline). The observed effects were statistically heterogeneous Test of heterogeneity (Q = 33.15 with df = 7, p <0.0001).

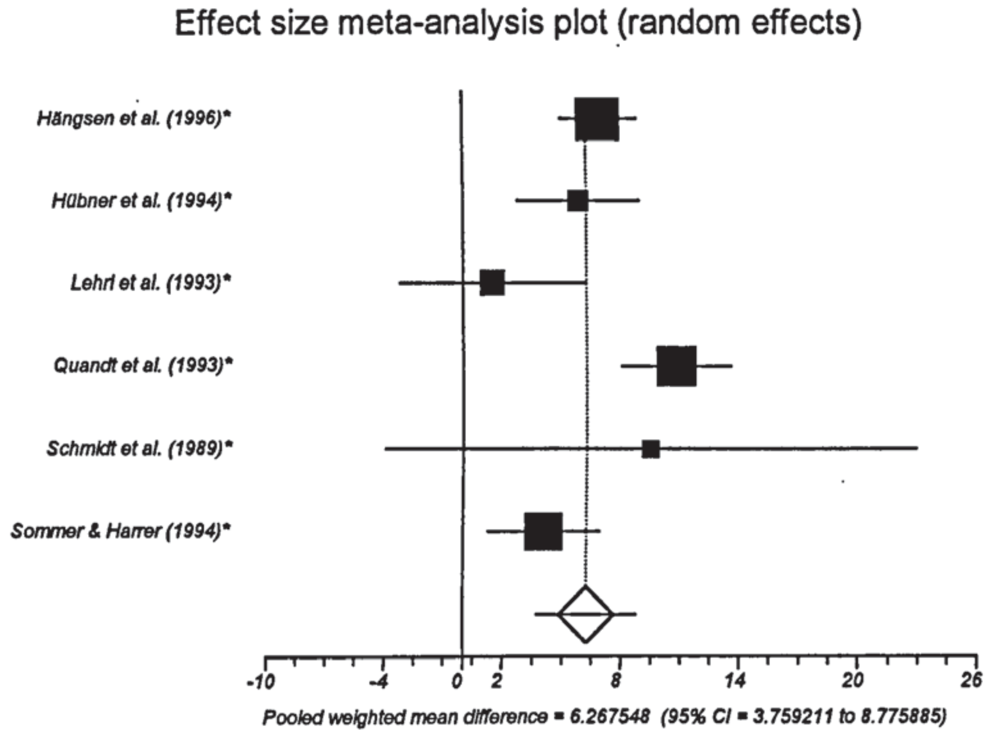
The squares give the point estimates and the horizontal line across each point gives the 95% confidence interval. The size of the square represents the weight assigned to the study concerned. The pooled estimate is shown as a diamond shape. The vertical dotted line is the pooled estimate of effect.



* Data derived from endpoint values

Figure 7.6 Subgroup analysis of trials with diagnostic classification (DSM-IIIR, DSM-IV, ICD-10) (change-from-baseline). The observed effects were statistically heterogeneous. Test of heterogeneity (Q = 33.15 with df = 7, p <0.0001).

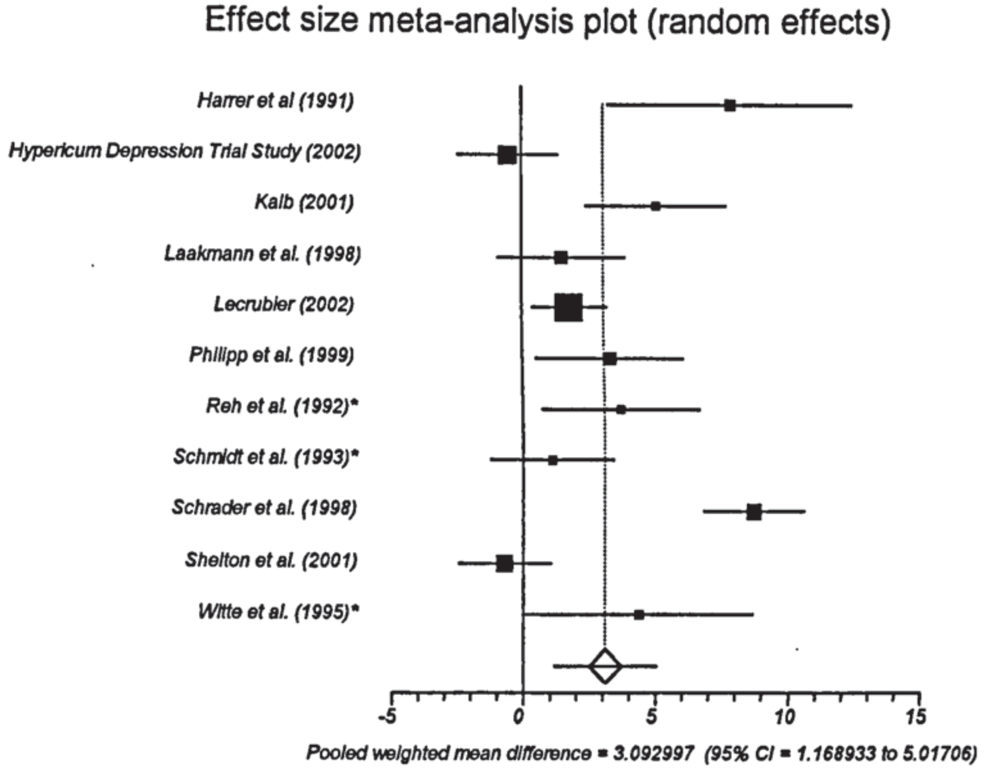
The squares give the point estimates and the horizontal line across each point gives the 95% confidence interval. The size of the square represents the weight assigned to the study concerned. The pooled estimate is shown as a diamond shape. The vertical dotted line is the pooled estimate of effect.



*Data derived from endpoint values

Figure 7.7 Subgroup analysis of trials lasting 4 weeks and less (change-from-baseline). The observed effects were statistically heterogeneous. Test of heterogeneity (Q = 90.08 with df = 8, p <0.0001).

The squares give the point estimates and the horizontal line across each point gives the 95% confidence interval. The size of the square represents the weight assigned to the study concerned. The pooled estimate is shown as a diamond shape. The vertical dotted line is the pooled estimate of effect.



* Data derived from endpoint values

Figure 7.8 Subgroup analysis of trials lasting more than 4 weeks (change-from-baseline). The observed effects were statistically heterogeneous. Test of heterogeneity (Q = 53.68 with df = 10, p <0.0001).

The squares give the point estimates and the horizontal line across each point gives the 95% confidence interval. The size of the square represents the weight assigned to the study concerned. The pooled estimate is shown as a diamond shape. The vertical dotted line is the pooled estimate of effect.

Table 7.1 Summary of subgroup analyses in placebo-controlled trials

Subgroup	Pooled RR	Pooled WMD
Diagnostic Classification		
ICD-9	2.82 (2.17 - 3.66)	5.13 (2.43 - 7.83)
DSM-IIIR, DSMIV, ICD-10	1.59 (1.27 - 2.00)	3.34 (1.03 - 5.64)
	*U = 11.00, p =0.009	U = 24.50, p = 0.268
Length of Trials		
4 weeks and less	2.74 (1.91-3.95)	6.27 (3.76 – 8.78)
More than 4 weeks	1.58 (1.29-1.94)	3.09 (1.17 – 5.02)
	*U = 11.00, p = 0.010	U = 15.50, p = 0.078

* Significant difference with Mann-Whitney test

Despite the observed non-homogeneity of effects, within the subgroups, analyses revealed that the treatment effects were greater in trials using the early diagnostic classification (ICD-9) than those using the recent diagnostic classifications (DSM-IIIR, DSM-IV, ICD-10)(Table 7.1, Figure 7.9 and Figure 7.10). Length of trials also appeared to have an influence on treatment effects. A larger pooled RR was observed for the group of trials with shorter trial duration (4 weeks and less) than with trials lasting more than 4 weeks. A Mann-Whitney U test showed that while the difference was statistically significant for the analysis using response data, it was not significant with the data using change of HAMD score (Table 7.1).

Summary of subgroup meta-analyses (response rate)

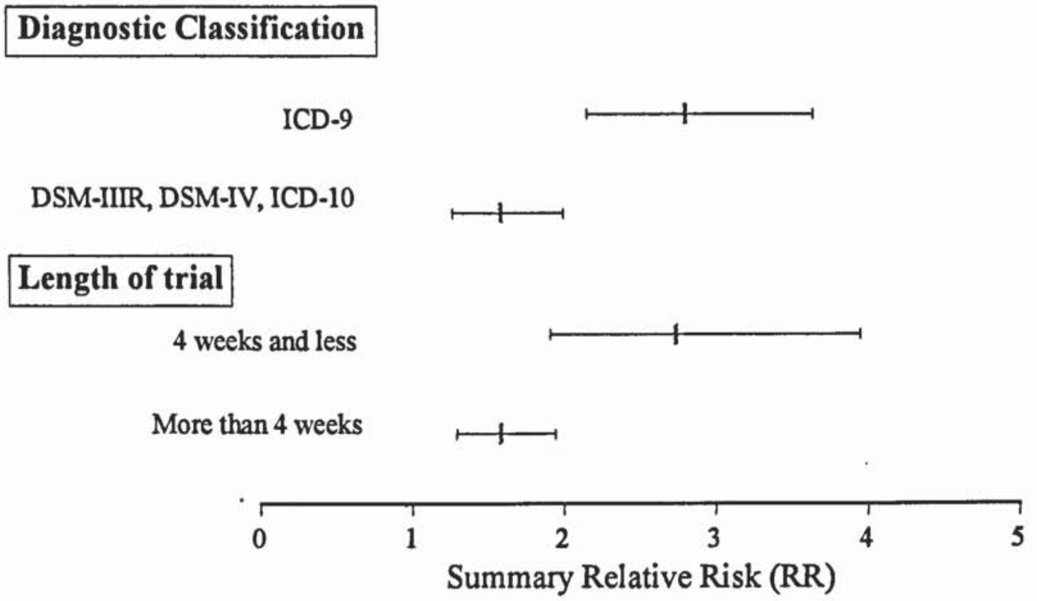


Figure 7.9 A vertical line across the horizontal bar indicates the point estimate of the summary RR for each subgroup of studies with the particular characteristics shown. The width of the horizontal line represents the 95% confidence interval of the summary RR for each subgroup. RR values of >1.0 represents an advantage for St. John's wort compared to placebo (all four estimates show superiority of St. John's wort).

Summary of subgroup meta-analyses (change-from-baseline)

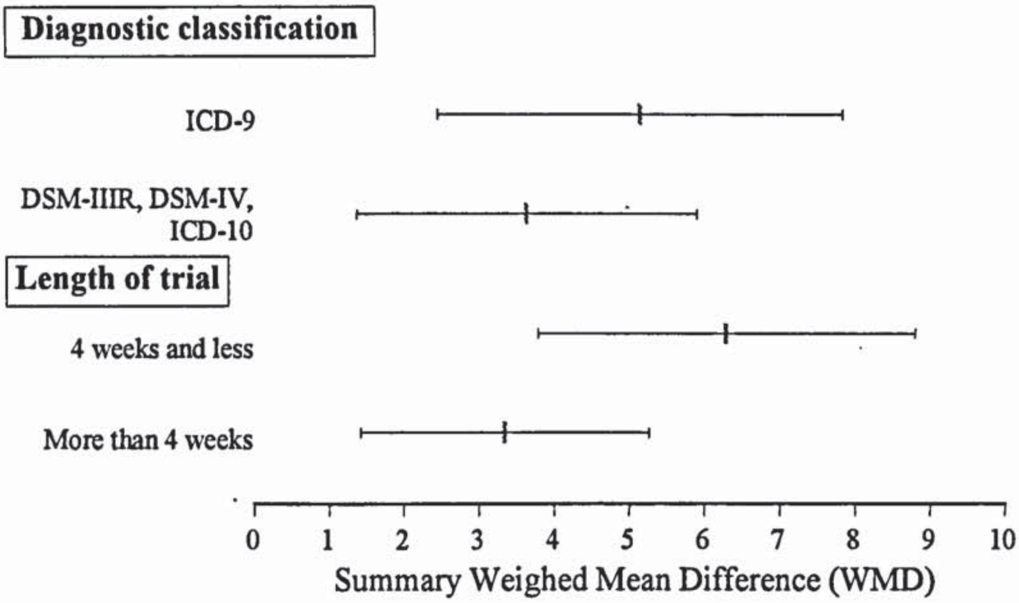


Figure 7.10 A vertical line across the horizontal bar indicates the point estimate of the summary WMD for each subgroup of studies with the particular characteristic shown. The width of the horizontal line represents the 95% confidence interval of the summary WMD for each subgroup.

7.3.2 Loglinear Analysis

The log linear final model suggested that only two-way interactions were significant. The association was between diagnostic classification and length of trials (χ^2 8.917, p = 0.0028).

7.4 Discussion

Subgroup analyses

Despite evidence of the effectiveness of St. John's wort, there is substantial heterogeneity among trials comparing St. John's wort with placebo for depression.

Most of the early trials used the ICD-9 classifications of depression. Additionally these trials were mostly of four-week duration. Treatment effects observed were greater in trials using ICD-9 classification and shorter trial length. Therefore, it seems that treatment effects were greater in older trials. More recent trials have not replicated the superiority shown by St. John's wort over placebo in those earlier trials.

The differences observed in subgroup analyses using WMD scores were not statistically significant, unlike the analyses using RR. This occurrence could be explained by the fact that categorizing data into responders and non-responders based on a response of a 50% improvement in HAMD scores, may inflate differences between groups if data were clustered around the point of cut-off (Moncrieff, 2001).

It is highly likely that there are several factors, which interact with each other to cause heterogeneity of treatment effects. We have not explored other variables, which may have contributed, to the heterogeneity of effects. Year in which the trial was conducted, the trial quality, and types of St. John's wort preparations, are potentially influential variables.

Trial quality was not examined because of the difficulty to adopt a reliable method to assess the trial quality. Quality rating is mostly based on the report of the study, which often does not give an accurate assessment of some elements of quality (Huwiler-Muntener et al., 2002). There is still considerable debate on how to assess trial quality (Moher et al., 1996; Juni et al., 2001). Since several different preparations (with various dosage forms) from ten different manufacturers were involved in the trials, subgroup analyses according to types of preparations were not possible. In addition, the subgroup analyses on types of St. John's wort preparations by the Cochrane Review (Linde and Mulrow, 2003) did not show clear evidence that any one preparation is better than the other.

Subgroup analyses sometimes referred to as data-dredging are subject to many recognised limitations, including false associations (Oxman and Guyatt, 1992). Individual patient data meta-analysis may overcome this problem (Lau et al., 1997; Stewart and Parmar, 1993; Stewart and Clarke, 1995). Although this approach is potentially powerful, it was not possible in our study because of lack of access to the primary data.

Generally subgroup analyses are post-hoc analyses, more useful for generating hypotheses than for testing them (Song, 1999).

Loglinear Analysis

Because multiple trial characteristics may actually be correlated, estimation of the influence of different predictor variables on treatment effect in subgroup analysis may be unreliable. We used loglinear analysis to examine the correlation between the characteristics of the trials. Although the analysis has severe limitations in our study given the small set of 18 trials, our results suggest confounding between the variables, length of trial and diagnostic classification.

7.5 Conclusions

Our analysis provides some insights into what the sources of heterogeneity in treatment effects might be. The subgroup analyses show that different diagnostic classifications and different length of trials are possible sources of heterogeneity. The loglinear analysis revealed the confounding between the variables, diagnostic classification and length of trial.

CHAPTER 8

General discussion

The first part of the study was a survey carried out in a teaching hospital in Malaysia with the aim of identifying variables that are predictive of physicians' recommending herbal medicines to patients. Given that conventionally trained physicians in Malaysia are relatively unfamiliar with herbal medicine, which is not widely taught in medical schools, the finding that about 19% of physicians recommended herbal medicines was quite unexpected. This figure is considerably higher than the 2.0% reported in a New Zealand study (Marshall et al., 1990) and the 3.6% reported in a US study (Jump et al., 1998), but much lower than the 78% reported in a German study (Himmel et al., 1993). As demonstrated by Danesi (1993) culture influences perceptions of health and disease. Therefore, the variations observed probably reflect the cultural differences in the beliefs and attitudes of the physicians towards herbal medicines.

Use of logistic regression modelling to examine influential variables in relation to recommending herbal medicines has the advantage that the normal distributional assumptions of standard linear regression can be overcome by the use of the logit link function [log p/1-p]. As with any linear regression model, interdependency between predictor variables leads to incorrect estimation of the coefficients (Dohoo et al., 1996). This problem is termed multicollinearity. Two general approaches have been suggested to deal with multicollinearity. The first approach involves excluding variables, which are associated. The second approach involves creating scores, which combine data from multiple variables into a single variable. While correlation analysis suggested some correlation between the variables INTEREST and PERSONAL USE, this was only moderate [r = 0.49; p =0.01 (2-tailed)] and both variables were retained in the model.

We hypothesised that physicians who were interested in receiving training were likely to recommend herbal medicines to their patients, as reported by Berman et al. (1998) and Verhoef and Sutherland (1995). The results were consistent with this hypothesis. That was not surprising. What was surprising was the magnitude of the effect. Physicians who were interested in receiving training were 15 times as likely to recommend herbal medicines as those who were not interested. We speculate that, given the widespread use of herbal medicines in Malaysia, those physicians wanted the training to ensure appropriate and scientifically-based recommendations.

The most prominent limitation of this survey is its likely limited generalisability to other time periods and other settings. The physicians we surveyed, were based in a teaching hospital and may not be representative of physicians based in other settings such as non-teaching hospitals, hospital in rural areas and private practices.

Given the popularity of herbal medicines and the inadequate regulation of such medicines in Malaysia, our findings provide baseline information for future studies. Additionally, our study highlights the need to include training in herbal medicine in the medical curriculum so that physicians can make more informed evidence-based decisions.

The second part of our study concerns a critical evaluation of the evidence for St. John's wort for depression. In contrast to other herbal medicines, the efficacy of St. John's wort for depression has been widely investigated in controlled trials (Linde et al., 1996). The available evidence is considered to be according to USP criteria (USP Press Release, 2000). Our results are generally in agreement with those of the Cochrane meta-analysis (Linde and Mulrow, 2003). We found that St. John's wort was 1.9 times as likely as placebo to improve depression compared to almost the 2.5 times reported by the Cochrane reviewers. Our meta-analysis included two recent methodologically rigorous trials (Shelton et al., 2001; Hypericum Depression Trial Study Group, 2002), which did not support the claim that St. John's wort was better than placebo in depression. An important question was why the data from these two trials contradicted the results from 16 trials that had shown St. John's wort to be effective. Since almost all the trials included in the meta-analysis have had serious methodological flaws (Linde and Mulrow, 2003), it was difficult to determine if the differences in outcome were attributable to chance, to methodological inadequacies, or to systematic differences in study characteristics.

Additionally, the lack of any published negative studies of St. John's wort suggests publication bias. This is not surprising because it is difficult to publish data demonstrating absence of effect (Easterbrook et al., 1991). Thus, the RR 1.9 estimated by us may overestimate the true effect of St. John's wort.

There were insufficient trials comparing St. John's wort with conventional antidepressants. Most of the available trials were non-inferiority trials unable to demonstrate assay sensitivity and to define the non-inferiority margin (Snapinn, 2000). Therefore we are unable to draw robust inferences about the equivalence in efficacy between St. John's wort and conventional antidepressants.

While the pooled response data indicates a substantial beneficial effect for St. John's wort over placebo in depression, the interpretation of this evidence is difficult mainly because of serious methodological flaws in the trials and inter-trial heterogeneity (Chapter 4). The sources of heterogeneity in these trials include: (a) use of many different preparations of St. John's wort which may have variable active constituents (Nahrstedt and Butterweck, 1997; Busse, 2000) (b) use of several diagnostic classification systems resulting in the inclusion of diagnostically heterogeneous patient-groups (c) use of different forms of instruments for the same outcome measures (HAMD rating scales with different numbers of items) (d) variable length of follow-up.

Given the nature of herbal products, blinding participants and researchers in clinical trials of herbal medicines is not easy and may introduce bias (Gaus and Hogel, 1995; Wong et al., 1998). In addition, like any other antidepressant trials, the subjective nature of outcome assessment also makes interpretation difficult. This subjective assessment could be a serious flaw because most included trials rely mainly on potentially biased clinician assessment (HAMD) rather than on patient-rated assessment (e.g. the BDI). Lambert et al. (1986) showed that patient-rated measures led to significantly smaller effect sizes than clinician-rated measures because patients tend to rate improvement lower compared to clinicians.

Individual patient meta-analysis is a potentially more reliable method for summarising the results of different trials (Lau et al., 1997; Stewart and Clarke, 1995; Stewart and Parmar, 1993). Although such an analysis is potentially powerful, it was not possible in our study because of lack of access to the primary data.

To examine the sources of heterogeneity in a meta-analysis, methods such as meta-regression and subgroup analysis have been used (Berlin and Antman, 1994; Thompson, 1994). However these methods sometimes referred to as data-dredging may lead to spurious conclusions (Oxman and Guyatt, 1992;

Thompson and Higgins, 2002) and are more appropriate for hypothesis generation. Because multiple trial characteristics may actually be correlated, estimation of the influence of different predictor variables on treatment effect in a meta-regression may be unreliable. This is especially true for St. John's wort trials, where there are few trials in the meta-analysis but many possible trial characteristics, and these characteristics can be highly correlated.

In this study we used loglinear analysis, to investigate potential correlations between the variables, which contribute to heterogeneity of effects. Correlation between diagnostic classifications, length of trials and treatment effects were examined. Despite the limitation of sample size (18 trials) results suggest that length of trials was associated with diagnostic classification.

Our analysis suggests that St. John's wort may exerts some beneficial effect on depression. However, serious questions remain regarding the research design of most of the studies analysed. Additionally the magnitude of this effect cannot be estimated with precision because of the heterogeneity in the published trials. Likewise, while the published studies suggest that the efficacy of St. John's wort is similar to that of conventional antidepressants, equivalence cannot be inferred. In the light of the findings from two recent large and well-done studies it is sensible to conclude that St. John's wort does not produce clinically meaningful responses in the treatment of depression.

Appendix 1. Questionnaire form

A QUESTIONNAIRE ON ATTITUDES AND PRACTICES REGARDING HERBAL MEDICINES

Dear Physicians,

We are carrying out a survey on attitudes and practices with respect to herbal medicines.

We would very much appreciate it if you could participate in this survey. All of the information you provide will be treated as completely confidential and it will not be possible for anyone to identify the information you have given. Thank you for your assistance.

PART A: **Attitudes toward Herbal Medicines**

Herbal medicines are defined as commercially packaged products, <u>usually</u> sold over the counter, that contain as their primary active ingredients materials extracted or derived from natural plant sources (e.g.: ginseng, gingko biloba, garlic... etc.)

1. In general, how would you rate your personal interest in herbal medicines? Please mark a point on the line below.

Not interested — Most interested

1 2 3 4 5

2. In general, do you perceive herbal medicines to be useful for treating patients? Please mark a point on the line below.

Not useful — Most useful

1 2 3 4 5

3. In general, do you perceive herbal medicines to be safe? Please mark a point on the line below.

Not safe — Most safe

1 2 3 4 5

4. For each of the statement below, please indicate the extent of your agreement or disagreement by placing a tick in the appropriate column.

Statement	Strongly agree	Agree	Neither agree or disagree	Disagree	Strongly disagree
	5	4	3	2	1
a. Herbal medicines should only be used in the treatment of minor health problems (e.g., common colds and coughs)					
b. Herbal medicines should NOT be used for serious health problems (e.g., chronic (e.g., chronic asthma, stroke)					
c. Herbal medicine should only be used if there is evidence of effectiveness from randomised clinical trials.					
d. Scientific evidence required for herbal medicines must be similar to that of conventional medicines.					

5. How essential is the following types of evidence in helping you to assess the beneficial and harmful effects of herbal medicines? Please indicate your response using the scale from 1 to 5, where **"5"** means **very essential** and **"1" least essential.**

Types of evidence					
Randomised Controlled Trials	1	2	3	4	5
Epidemiological Studies	1	2	3	4	5
Historical evidence of use	1	2	3	4	5
Published Case Studies	1	2	3	4	5
Success in own practice	1	2	3	4	5
Experts Opinion	1	2	3	4	5
Patient reports:	1	2	3	4	5

PART B: **Personal Belief and Practice of Herbal Medicine**

1. Do you use herbal medicines for yourself?

Yes	
No	

Please go to **Question 2**

If **"YES"**, please name any 3 herbal medicines that you use.

Herbal medicine A: ______________________________
Herbal medicine B: ______________________________
Herbal medicine C: ______________________________

a. Please tick the reason(s) for their use.

Reasons (s) for Use	Herbal medicine A	Herbal medicine B	Herbal medicine C
Prevention of disease			
Improved feeling of well-being			
Relieve symptoms			
Benefits to immune system			
Treatment of disease			
Increased energy			
Others: please specify			

b. Have you experienced beneficial effect(s) from the use of these herbal medicines?

Yes, for all three products	
Yes, for 2 products	
Yes, for 1 product	
No	
Not Sure	

c. Have you experienced harmful effect(s) from the use of these herbal medicines ?

Yes, for all three products	
Yes, for 2 products	
Yes, for 1 product	
No	
Not Sure	

2. Have your patients discussed herbal medicines with you?

Yes	
No	

3. Do you know whether your patients are taking herbal medicines?

Yes		
Never ask		Please go to **Question 4**
No		Please go to **Question 4**

If **"YES"**, please name any 3 herbal medicines/products used.

Herbal medicine A: ________________________________
Herbal medicine B: ________________________________
Herbal Medicine C: ________________________________

a. Please tick the reason(s) for their use.

Reasons (s) for Use	**Herbal medicine A**	**Herbal medicine B**	**Herbal medicine C**
Prevention of disease			
Improved feeling of well-being			
Relieve symptoms			
Benefits to immune system			
Treatment of disease			
Increased energy			
Others: please specify			

4. Have you recommended herbal medicines to your patients?

Yes		
No		Please go to **Question 5**

If **"YES"**, please name any 3 herbal medicines you have recommended.

i ______________________________
ii ______________________________
iii ______________________________

5. If your patients are taking herbal medicines, do you know who influenced him/her to take these medicines?

Yes		
No		Please go to **PART C**

If **"YES"**, please tick the source of the influence

Media sources (newspapers, TV, radio, and internet…)	
Friends/family/relatives	
Traditional/herbal practitioners	
Pharmacy/health food store	
Do not know	
Others (please specify)	

PART C: Training in Herbal Medicines

1. Have you had any training in herbal medicines?

Yes	
No	

2. Would you be interested in receiving training about herbal medicines?

Yes	
No	

3. Should education on herbal medicines be incorporated into standard medical curriculum at the undergraduate level?

Yes	
No	

4. Are you involved in herbal medicine research?

Yes	
No	

5. Do you think that there should be more research on effects of herbal medicines?

Yes	
No	

PART D: Demographic

1. Gender:

   Male □ Female □

2. Ethnic:

   Malay □ Chinese □

   Indian □ Others □
   (Please specify) ____________

3. Age (years)

   <25 □ 46-55 □

   26-35 □ >55 □

   36-45 □

4. Where did you receive your first degree/training?

Malaysia □ Overseas □

(Please specify the country) ____________

5. Where did you receive your higher degree/training?

Malaysia □ Overseas □
(Please specify the country) ____________

6. For how many years have you been practicing/lecturing? ____________

7. Are you directly involved in patient care?

Yes	
No	

8. Your category?

Consultant	
Lecturer	
Medical Officer	
House Officer	

9. Your Department? ______________________________

PART E: **Comments**

Do you have any general comments regarding herbal medicines and/or about this questionnaire?

Yes		Please give your comments below
No		

THANK YOU VERY MUCH FOR YOUR VALUABLE TIME.

Appendix 2. DSM-IIIR Classification: Criteria for depression

Symptoms

1. Depressed mood
2. Substantial weight loss or weight gain
3. Insomnia or hypersomnia
4. Feelings of worthlessness or inappropriate guilt
5. Recurrent thoughts of death or suicide or suicide attempt
6. Decreased interest or pleasure
7. Psychomotor retardation or agitation
8. Fatigue or loss of energy
9. Diminished ability to think or concentrate

The DSM-IIIR requires the presence of at least five of the symptoms listed above for a diagnosis of major depressive episode.

From *Diagnostic and Statistical Manual of Mental Disorders*, Third Edition

Appendix 3. DSM-IV Classification: Criteria for major depression

Symptoms

1. Depressed mood
2. Decreased interest or pleasure
3. Substantial weight loss or weight gain
4. Insomnia or hypersomnia
5. Feelings of worthlessness or inappropriate guilt
6. Recurrent thoughts of death or suicide or suicide attempt
7. Psychomotor retardation or agitation
8. Fatigue or loss of energy
9. Diminished ability to think or concentrate

The DSM-IV requires the presence over the last two weeks of at least five of the symptoms listed above and at least one of the symptoms (either 1 or 2) for a diagnosis of major depressive episode.

From *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition

Appendix 4. ICD-10 Classification: Criteria for major depression

The signs and symptoms of the Major Depressive disorder

For all 3 varieties (mild, moderate severe)

Core Symptoms

a. Depressed mood
b. Loss of interest and enjoyment
c. Reduced energy, often leading to an increased tendency to fatigue and diminished activity

Two out of three of these core symptoms must be present for a diagnosis.

Associated signs and symptoms

a. Reduced concentration and attention
b. Reduced self-esteem and self-confidence
c. Feelings of guilt and unworthiness
d. Bleak and pessimistic views of the future
e. Ideas or acts of self-harm or suicide
f. Disturbed sleep
g. Diminished appetite
h. Weight loss
i. Loss of libido

Additionally, at least one additional associated symptom for a total of four symptoms needs to be present for a diagnosis.

Assessment of the severity

Mild depressive disorder

At least two of the core symptoms and at least two of the associated symptoms must be present. The patient is usually distressed by these but will probably be able to continue with most activities

Moderate depressive disorder

More of the above symptoms are usually present and the patient is likely to have great difficulty in continuing with ordinary activities.

Severe depressive disorder

Several of the above symptoms are marked and distressing, typically loss of self-esteem, feeling of worthlessness or guilt. Suicidal thoughts and acts are common and a number of somatic symptoms are usually present.

From *WHO The International Statistical Classification of Diseases and Related Health problems.* Tenth Revision

Appendix 5. List of trials included

Behnke K, Jensen GS, Graubaum HJ, Gruenwald J. Hypericum Perforatum Versus Fluoxetine in the Treatment of Mild to Moderate Depression. *Advances in Therapy* 2002; 19(1): 43-52.

Bergmann R, Nubner J, Demling J. Behandlung leichter bis mittelschwerer Depressionen. *Therapiewoche Neurologie/Psychiatrie* 1993; 7: 235-40.

Brenner R, Azbel V, Madhusoodanan S, Pawlowska M. Comparison of an extract of hypericum (LI 160) and sertraline in the treatment of depression: a double-blind, randomized pilot study. *Clin Ther 2000*; 22(4): 411-9.

Halama P. Wirksamkeit des Johanniskrautextraktes LI 160 bei depressiver Verstimmung. *Nervenheilkunde* 1991; 10: 250-3.

Hängsen KD, Vesper J. Antidepressive Wirksamkeit eines hochdosierten Hypericum-Extraktes. *Muench Med Wschr* 1996; 138(3): 29-33.

Harrer G, Hubner WD, Podzuweit H. Effectiveness and tolerance of the hypericum extract LI 160 compared to maprotiline: a multicenter double-blind study. *J Geriatr Psychiatry Neurol* 1994; 7 Suppl 1: S24-S28.

Harrer G, Schmidt U, Kuhn U. "Alternative" Depressionsbehanddlung miteinem Hypericum-Extrakt. Therapiewoche Neurologie 1991; 5: 710-716.

Harrer G, Schmidt U, Kuhn U, Biller A. Comparison of equivalence between the St. John's wort extract LoHyp-57 and fluoxetine. *Arzneimittelforschung* 1999; 49(4): 289-96.

Hübner WD, Lande S, Podzuweit H. Hypericum treatment of mild depressions with somatic symptoms. *J Geriatr Psychiatry Neurol* 1994; 7 Suppl 1: S12-S14.

Hypericum Depression Trial Study Group. Effect of Hypericum perforatum (St John's wort) in major depressive disorder: a randomized controlled trial. *JAMA* 2002; 287(14): 1807-14.

Kalb R, Trautmann-Sponsel RD, Kieser M. Efficacy and tolerability of hypericum extract WS 5572 versus placebo in mildly to moderately depressed patients. A randomized double-blind multicenter clinical trial. *Pharmacopsychiatry* 2001; 34(3): 96-103.

Laakmann G, Schule C, Baghai T, Kieser M. St. John's wort in mild to moderate depression: the relevance of hyperforin for the clinical efficacy. *Pharmacopsychiatry* 1998; 31 Suppl 1: 54-9.

Lecrubier Y, Clerc G, Didi R, Kieser M. Efficacy of St. John's wort extract WS 5570 in major depression: a double-blind, placebo-controlled trial. *Am J Psychiatry* 2002; 159(8): 1361-6.

Lehrl S, Woelk H. Ergebnisse von Messungen der kognitiven Leistungsfähigkeit bei Patienten unter der Therapie mit Johanniskraut. *Nervenheilkunde* 1993; 12: 281-4.

Philipp M, Kohnen R, Hiller KO. Hypericum extract versus imipramine or placebo in patients with moderate depression: randomised multicentre study of treatment for eight weeks. *BMJ* 1999; 319(7224): 1534-8.

Quandt J, Schmidt U, Schenk N. Ambulante Behandlung leichter und mittelschwerer depressiver Verstimmungen. *Der Allgemeinarzt* 1993; 15(2): 97-102.

Reh C, Laux P, Schenk N. Hypericum-Extrakt bei Depressionen - eine wirksame Alternative. *Therapiewoche* 1992; 42: 1576-81.

Schlich D, Braukmann F, Schenk N. Behandlung depressiver Zustände mit Hypericinium. *Psycho* 1987; 13: 440-7.

Schmidt U, Schenk N, Schwarz I, Vorberg G. Zur Therapie depressiver Verstimmungen. *Psycho* 1989; 15: 665-71.

Schmidt U, Sommer H. Johanniskraut-Extrakt zur ambulanten Therapie der Depression. *Fortschritte Medizin* 1993; 111: 339-42.

Schrader E. Equivalence of St John's wort extract (Ze 117) and fluoxetine: a randomized, controlled study in mild-moderate depression. *Int Clin Psychopharmacol* 2000; 15(2): 61-8.

Schrader E, Meier B, Brattström A. Hypericum treatment of mild-moderate depression in a placebo-controlled study. A prospective, double-blind, randomized, placebo-controlled, multicentre study. *Hum Psychopharmacol* 1998; 13: 163-9.

Shelton RC, Keller MB, Gelenberg A *et al.* Effectiveness of St John's wort in major depression: a randomized controlled trial. *JAMA* 2001; 285(15): 1978-86.

Sommer H, Harrer G. Placebo-controlled double-blind study examining the effectiveness of an hypericum preparation in 105 mildly depressed patients. *J Geriatr Psychiatry Neurol* 1994; 7 Suppl 1: S9-11.

van Gurp G, Meterissian GB, Haiek LN, McCusker J, Bellavance F. St John's wort or sertraline? Randomized controlled trial in primary care. *Canadian Family Physician* 2002; 48: 905-12.

Vorbach EU, Arnoldt KH, Hubner WD. Efficacy and tolerability of St. John's wort extract LI 160 versus imipramine in patients with severe depressive episodes according to ICD-10. *Pharmacopsychiatry* 1997; 30 Suppl 2: 81-5.

Vorbach EU, Hubner WD, Arnoldt KH. Effectiveness and tolerance of the hypericum extract LI 160 in comparison with imipramine: randomized double-blind study with 135 outpatients. *J Geriatr Psychiatry Neurol* 1994; 7 Suppl 1: S19-S23.

Wheatley D. LI 160, an extract of St. John's wort, versus amitriptyline in mildly to moderately depressed outpatients--a controlled 6-week clinical trial. *Pharmacopsychiatry* 1997; 30 Suppl 2: 77-80.

Witte B, Harrer G, Kaplan T, Podzuweit H, Schmidt U. Behandlung depressiver Verstimmungen mit einem hochkonzentrierten Hypericumpräparat - eine multizentrische plazebokontrollierte Doppelblindstudie. *Fortschr Med* 1995; 113: 404-8.

Woelk H. Comparison of St John's wort and imipramine for treating depression: randomised controlled trial. *BMJ* 2000; 321(7260): 536-9.

Appendix 6. List of trials excluded and reasons for exclusion

Reference of Study	Reasons for exclusion
Albertini H. In: Boiron J, Belon P, Hariveau E, editor(s). Recherche en homeopathie. Lyon: Fondation francaise pour la recheche en homeopathie, 1986: 75-77.	RCT of St. John's wort in homeopathic preparation for dental neuralgia
Bendre VV, Dharmadhikari SD. Arnica montana and hypericum in dental practice. Hahnemann Gleanings 1980; 47: 70-72.	Homeopathic preparations of St. John's wort in dental practice
Bernhardt M, Liske E, Ebeling L. Hypericum perforatum in der Therapie leichter bis mittelschwerer Depressionen: Vergleich der antidepressiven Wirksamkeit von zwei unterschiedlichen Dosierungsschemata. Bonn, 5. Phytotherapie-Kongreß 1993	RCT comparing two doses of St. John's wort for depression
Brockmöller J, Reum T, Bauer S, Kerb R, Hübner WD, Roots I. Hypericin and pseudohypericin: pharmacokinetics and effects on photosensitivity in human. Pharmacopsychiatry 1997; 30(suppl.2): 94-101.	RCT on pharmacokinetics and photosensitivity in healthy volunteers
Ditzler K, Gessner B, Schatton WFH, Willems M. Clinical trial on Neuropas versus placebo in patients with mild to moderate depressive symptoms: a placebo-controlled, randomised double-blind study. Complementary Therapies in Medicine 1994; 2: 5-13.	RCT on combination preparations of St. John's wort
Hoffmann J, Kühl ED. Therapie von depressiven Zuständen mit Hypericin. Z Allgemeinmed 1979; 55: 776-782.	Only available as abstract. Degree of depression not clear
Hoffmann J, Kühl ED. Therapie von depressiven Zuständen mit Hypericin. Z Allgemeinmed 1979; 55: 776-782.	Outcome measured was unclear

Reference of Study	**Reasons for exclusion**
Hottenrott K, Sommer HM, Lehrl S, Hauer H. Der Einfluss von Vitamin E und Johanniskraut-Trockenextrakt auf die Ausdauerleistungsfähigkeit von Wettkampfsportlern. Eine placebo-kontrollierte Doppelblindstudie mit Langstreckenläufern und Triathleten. Deutsche Zeitschrift für Sportmedizin 1997; 48: 22-27.	RCT on St. John's wort in combination with Vitamin E to enhance performance of athletes
Johnson D. Neurophysiologische Wirkungen von Hypericum im Doppelblindversuch mit Probanden. Nervenheilkunde 1991; 10: 316-317.	RCT on neurophysiological effects of St. John's wort.
Johnson D, Ksciuk H, Woelk H, Sauerwein-Giese E, Frauendorf A. Effetcs of hypericum extract LI 160 compared with maprotiline on resting EEG and evoked potentials in 24 volunteers. J Geriatr Psychiatry Neurol 1994; 7(suppl 1): S44-46.	RCT comparing neurophysiological effects of St. John's wort and maprotiline
Kniebel R, Burchard JM. Zur Therapie depressiver Verstimmungen in der Praxis. Z Allgemeinmed 1988; 64: 689-696.	RCT on St. John's wort in combination with Valerian
König CD. Hypericum perforatum L. (gemeines Johanniskraut) als Therapeutikum bei depressiven Verstimmungszuständen - eine Alternative zu synthetischen Arzneimitteln. University of Basel: Thesis, 1993	Outcome not measured by HAMD Rating scale
Kugler J, Schmidt A, Groll S, Weidenhammer W. Zur Pharmakodynamik eines Hypericum-Extraktes. Untersuchungen bei Patienten mit depressiven Zuständen im Vergleich zu Bromazepam und Placebo. Ztschr Allgemeinmed 1990; 66 (suppl): 13-20.	RCT on pharmacodynamics of St. John's wort
Lenoir S, Degenring FH, Saller R. A double-blind randomised trial to investigate three different concentrations of a standardised fresh plant extract obtained from the shoot tips of Hypericum perforatum L. Phytomedicine. 1999 Jul; 6(3): 141-6.	RCT comparing three different concentrations of St. John's wort for depression

Reference of Study	**Reasons for exclusion**
Maisenbacher J, Schmidt U, Schenk. N. Therapie mit Hypericum bei Angstzuständen. Therapiewoche Neurologie Psychiatrie 1995; 9: 65-60.	RCT of St. John's wort for anxiety
Martinez B, Kasper S, Ruhrmann B, Möller HJ. Hypericum in the treatment of seasonal affective disorders. J Geriatr Psychiatry Neurol 1994; 7(suppl 1): S29-33.	RCT of St. John's wort in patients with seasonal affective disorder
Panijel J. Die Behandlung mittelschwerer Angstzustände. Therapiewoche 1985; 41: 4659-4668.	RCT of St. John's wort in combination with Valerian for anxiety.
Osterheider M, Schmidtke A, Beckmann H. Behandlung depressiver Syndrome mit Hypericum (Johanniskraut - eine placebokontrollierte Doppelblindstudie. Fortschr Neurol Psychiatr 1992; 60(suppl.2): 210-211.	Published as abstract only.
Schmidt U, Harrer G, Kuhn U, Berger-Deinert W, Luther D. Wechselwirkungen von Hypericin-Extrakt mit Alkohol. Nervenheilkunde 1993; 12: 314-319	RCT on interactions of St. John's wort and alcohol in healthy volunteers
Schulz H, Jobert M. Effects of hypericum extract on the sleep EEG in older volunteers. J Geriatr Psychiatry Neurol 1994; 7(suppl 1): S39-43.	RCT on the effects of St. John's wort on the sleep EEG.
Spielberger E. Johanniskraut-Präparat lindert selbst mittelschwere Depressionen. Ärztliche Praxis 1985; 37: 2546-2547.	RCT comparing two St. John's wort preparations for depression
Steger W. Depressive Verstimmungen. Ztschr Allgemeinmed 1985; 61: 914-918.	RCT of St. John's wort in combination with Valerian.
Staffeldt B, Kerb R, Brockmöller J, Ploch M, Roots I. Pharmacokinetics of hypericin and pseudohypericin after oral intake of the hypericum perforatum extract LI 160 in healthy volunteers. J Geriatr Psychiatry Neurol 1994;	RCT on pharmacokinetics of St. John's wort.

Reference of Study	**Reasons for exclusion**
7(suppl 1): S47-53.	
Vesper J, Ploch M. Multi-center double blind study examinng the antidepressant effectiveness of the hypericum extract LI 160. J Geriatr Psychiatry Neurol 1994; 7(suppl 1): S15-18.	Trial first published with 72 patients and republished in 1996 with 108 patients. (Same study as Hangsen et al. 1996)
Warnecke G. Beeinflussung klimakterischer Depressionen. Ztschr Allgemeinmed 1986; 62: 1111-1113.	Diazepam used as comparator.
Werth W. Psychotonin M versus Imipramin in der Chirurgie. Der Kassenarzt 1989; 15: 64-68.	Single blind RCT
Wienert V, Classen R, Hiller KO. Zur Frage der Photosensibilisierung von Hypericin in einer Baldrian-Johanniskraut-Kombination - klinisch-experimentelle, plazebokontrollierte Vergleichsstudie. Lübeck-Travemünde, 3rd Phytotherapy Congress 1991.	RCT on photosensitivity reaction after application of a combination of St. John's wort and Valerian

Appendix 7. Statistics used to express the effect of an intervention

Table 1. Statistical Measures Used to Express the Effect of an Intervention
(*From:* Moher: Arch Pediatr Adolesc Med, Volume 152(9). September 1998: 915-920)

References

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders DSM-IV Fourth Edition 1994; Washington, DC: APA.

Astin JA. Why patients use alternative medicine: results of a national study. *JAMA* 1998; 279(19): 1548-53.

Astin JA, Marie A, Pelletier KR, Hansen E, Haskell WL. A review of the incorporation of complementary and alternative medicine by mainstream physicians. *Arch Intern Med* 1998; 158(21): 2303-10.

Barrett B, Kiefer D, Rabago D. Assessing the risks and benefits of herbal medicine: an overview of scientific evidence. *Altern Ther Health Med.* 1999; 5(4): 40-9.

Basoglu M, Marks I, Livanou M, Swinson R. Double-blindness procedures, rater blindness, and ratings of outcome. Observations from a controlled trial. *Arch Gen Psychiatry* 1997; 54(8): 744-8.

Bausell RB, Lee WL, Berman BM. Demographic and health-related correlates to visits to complementary and alternative medical providers. *Medical Care* 2001; 39(2): 190-6.

Beck AT, Rush AJ, Shaw BF, Emery G. Cognitive therapy for depression. New York, Guilford, 1979.

Behnke K, Jensen GS, Graubaum HJ, Gruenwald J. Hypericum Perforatum Versus Fluoxetine in the Treatment of Mild to Moderate Depression. *Advances in Therapy* 2002; 19(1): 43-52.

Bensoussan A, Talley NJ, Hing Meal. Treatment of irritable bowel syndrome with Chinese herbal medicine. A randomized controlled study. *JAMA* 1998; 280: 1585-9.

Bergmann R, Nubner J, Demling J. Behandlung leichter bis mittelschwerer Depressionen. *Therapiewoche Neurologie/Psychiatrie* 1993; 7: 235-40.

Berlin JA, Antman EM. Advantages and limitations of metaanalytic regressions of clinical trial data. *Online J Curr Clin Trials* 1994; Doc No 134.

Berlin JA, Laird NM, Sacks HS, Chalmers TC. A comparison of statistical methods for combining event rates from clinical trials. *Stat Med* 1989; 8(2): 141-51.

Berman BM, Bausell RB, Lee WL. Use and referral patterns for 22 complementary and alternative medical therapies by members of the American College of Rheumatology: results of a national survey. *Arch Intern Med* 2002; 162(7): 766-70.

Berman BM, Singh BB, Hartnoll SM, Singh BK, Reilly D. Primary care physicians and complementary-alternative medicine: training, attitudes, and practice patterns. *J Am Board Fam Pract* 1998; 11(4): 272-81.

Berman BM, Singh BK, Lao L, Singh BB, Ferentz KS, Hartnoll SM. Physicians' attitudes toward complementary or alternative medicine: a regional survey. *J Am Board Fam Pract* 1995; 8(5): 361-6.

Blumberg DL, Grant WD, Hendricks SR, Kamps CA, Dewan MJ. The physician and unconventional medicine. *Altern Ther Health Med* 1995; 1(3): 31-5.

Borkan J, Neher JO, Anson O, Smoker B. Referrals for alternative therapies. *J Fam Pract* 1994; 39(6): 545-50.

Brenner R, Azbel V, Madhusoodanan S, Pawlowska M. Comparison of an extract of hypericum (LI 160) and sertraline in the treatment of depression: a double-blind, randomized pilot study. *Clin Ther* 2000; 22(4): 411-9.

Brevoort P. The booming U.S. botanical market - A new overview. *Herbalgram* 1998; 44: 33-46.

Buckman R, Lewith G. What does homoeopathy do--and how? *BMJ* 1994; 309(6947): 103-6.

Burg MA, Kosch SG, Neims AH, Stoller EP. Personal use of alternative medicine therapies by health science center faculty. *JAMA* 1998; 280(18): 1563.

Busse W. The significance of quality for efficacy and safety of herbal medicinal products. *Drug Inf J* 2000; 34: 15-23.

Carroll BJ, Feinberg M, Smouse PE, Rawson SG, Greden JF. The Carroll rating scale for depression. I. Development, reliability and validation. *Br J Psychiatry* 1981; 138:194-200

Casey P. Adult Adjustment Disorder: A Review of Its Current Diagnostic Status. *Journal of Psychiatric Practice* 2001; 7(1): 32-40.

Chatterjee SS, Bhattacharya SK, Wonnemann M, Singer A, Muller WE. Hyperforin as a possible antidepressant component of hypericum extracts. *Life Sci* 1998; 63(6): 499-510.

Cherniack EP, Senzel RS, Pan CX. Correlates of use of alternative medicine by the elderly in an urban population. *J Altern Complement Med* 2001; 7(3): 277-80.

Colditz GA, Burdick E, Mosteller F. Heterogeneity in meta-analysis of data from epidemiologic studies: a commentary. *Am J Epidemiol* 1995; 142(4): 371-82.

Corbin-Winslow L, Shapiro H. Physicians want education about complementary and alternative medicine to enhance communication with their patients. *Arch Intern Med* 2002; 162(10): 1176-81.

Cui J, Garle M, Eneroth P, Bjorkhem I. What do commercial ginseng preparations contain? *Lancet* 1994; 344(8915): 134.

Danesi M. The semiotic representation of 'health' and 'disease'. In *Health and Cultures Exploring the Relationships*. Vol. 1. Oakville: Mosaic Press, 1993.

Depression Guideline Panel. Depression in primary care. Treatment of Major Depression. Clinical Practice Guideline. Rockland, Md: U.S. Department of Health and Human Services, Agency for Health Care Policy and Research, 1993.

Dohoo IR, Ducrot C, Fourichon C, Donald A, Hurnik D. An overview of techniques for dealing with large numbers of independent variables in epidemiologic studies. *Preventive Veterinary Medicine 1997*; 29(3): 221-39.

Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997; 315(7109): 629-34.

Egger M, Smith GD, Phillips AN. Meta-analysis: principles and procedures. *BMJ* 1997b; 315(7121): 1533-7.

Eisenberg DM, Davis R, Ettner SL, Appel S, Wilkey S, Van Rompay M, Kessler RC. Trends in Alternative Medicine Use in the United States, 1990-1997: results of a Follow-up National Survey. JAMA 1998; 280(18): 1569-1575.

Eisenberg DM, Kessler RC, Foster C, Norlock FE, Calkins DR, Delbanco TL. Unconventional medicine in the United States. Prevalence, costs, and patterns of use. *N Engl J Med* 1993; 328(4):246-52

Endicott J, Cohen J, Nee J, Fleiss J, Sarantakos S. Hamilton Depression Rating Scale. Extracted from Regular and Change Versions of the Schedule for Affective Disorders and Schizophrenia. *Arch Gen Psychiatry* 1981; 38(1): 98-103.

Ernst E, Resch KL, White AR. Complementary medicine. What physicians think of it: a meta-analysis. *Arch Intern Med* 1995; 155(22): 2405-8.

Eskinazi DP. Factors that shape alternative medicine. JAMA. 1998 Nov 11; 280(18): 1621-3.

Faravelli C, Albanesi G, Poli E. Assessment of depression: a comparison of rating scales. *J Affect Disord* 1986; 11(3): 245-53.

Fleiss JL. Measures of effect size for categorical data. New York: Russell Sage Foundation, 1993: 245-81.

Follmann D, Elliott P, Suh I, Cutler J. Variance imputation for overviews of clinical trials with continuous response. *J Clin Epidemiol* 1992; 45(7): 769-73.

Freemantle N, Anderson IM, Young P. Predictive value of pharmacological activity for the relative efficacy of antidepressant drugs. Meta-regression analysis. *Br J Psychiatry* 2000; 177: 292-302.

Gaster B, Holroyd J. St John's Wort for depression: a systematic review. *J. Arch Intern Med* 2000; 160: 152-6.

Gaus W, Hogel J. Studies on the efficacy of unconventional therapies. Problems and designs. *Arzneimittelforschung* 1995; 45(1): 88-92.

Geddes J, Butler R. Depressive disorders. *Clin Evid* 2002; 7: 867-82.

Gelber RD, Goldhirsch A. The evaluation of subsets in meta-analysis. *Stat Med* 1987; 6: 371-88.

Gilbert G. Modelling Society. In: *An introduction to loglinear analysis for social researchers*. London: George Allen and Unwin, 1981.

Halama P. Wirksamkeit des Johanniskrautextraktes LI 160 bei depressiver Verstimmung. *Nervenheilkunde* 1991; 10: 250-3.

Hamilton M. Comparative value of rating scales. *Br J Clin Pharmacol* 1976; 3(1 Suppl 1): 58-60.

Hamilton M. Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol* 1967; 6(4): 278-96.

Hamilton M. Rating depressive patients. *J Clin Psychiatry* 1980; 41(12 Pt 2): 21-4.

Hamilton M. A rating scale for depression. *J. Neurol. Neurosurg., Psychiat* 1960; 23: 56-62.

Hängsen KD, Vesper J. Antidepressive Wirksamkeit eines hochdosierten Hypericum-Extraktes. *Muench Med Wschr* 1996; 138(3): 29-33.

Harrer G, Hubner WD, Podzuweit H. Effectiveness and tolerance of the hypericum extract LI 160 compared to maprotiline: a multicenter double-blind study. *J Geriatr Psychiatry Neurol* 1994; 7 Suppl 1: S24-S28.

Harrer G, Schmidt U, Kuhn U, Biller A. Comparison of equivalence between the St. John's wort extract LoHyp-57 and fluoxetine. *Arzneimittelforschung* 1999; 49(4): 289-96.

Hedlund J, Viewweg BW. The Hamilton Scale for Depression: a comparative review. *Journal of Operational Psychiatry* 1979; 10: 149-65.

Himmel W, Schulte M, Kochen MM. Complementary medicine: are patients' expectations being met by their general practitioners? *Br J Gen Pract* 1993; 43(371): 232-5.

Hübner WD, Lande S, Podzuweit H. Hypericum treatment of mild depressions with somatic symptoms. *J Geriatr Psychiatry Neurol* 1994; 7 Suppl 1:12-S14.

Hope K. A hidden problem: identifying depression in older people. Br J Community Nurs. 2003; 8(7): 314-20.

Hughes JR, O_Hara MW, Rehm LP. Measurement of depression in clinical trials: an overview. *J Clin Psychiatry* 1982; 43(3): 85-8.

Huwiler-Muntener K, Juni P, Junker C, Egger M. Quality of reporting of randomized trials as a measure of methodologic quality. *JAMA* 2002; 287(21): 2801-4.

Hypericum Depression Trial Study Group. Effect of Hypericum perforatum (St John's wort) in major depressive disorder: a randomized controlled trial. *JAMA* 2002; 287(14): 1807-14.

Ioannidis JP, Cappelleri JC, Lau J. Issues in comparisons between meta-analyses and large trials. *JAMA* 1998; 279(14): 1089-93.

Ioannidis JP, Lau J. Can quality of clinical trials and meta-analyses be quantified? *Lancet* 1998; 352: 590-1.

Jensen AG, Hansen SH, Nielsen EO. Adhyperforin as a contributor to the effect of Hypericum perforatum L. in biochemical models of antidepressant activity. *Life Sciences* 2001; 68(14): 1593-605.

Jump J, Yarbrough L, Kilpatrick S, Cable T. Physicians' Attitudes Toward Complementary and Alternative Medicine. *Integrative Medicine* 1998; 1(4): 149-53.

Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ* 2001; 323(7303): 42-6.

Kalb R, Trautmann_Sponsel RD, Kieser M. Efficacy and tolerability of hypericum extract WS 5572 versus placebo in mildly to moderately depressed patients. A randomized double-blind multicenter clinical trial. *Pharmacopsychiatry* 2001; 34(3): 96-103.

Katon W, Schulberg H. Epidemiology of depression in primary care. *Gen Hosp Psychiatry* 1992; 14(4): 237-47.

Kessler RC, McGonagle KA, Nelson CB, Hughes M, Swartz M, Blazer DG. Sex and depression in the national comorbidity survey. II: Cohort effects. *J Affect Disorders* 1994; 30(1): 15-26.

Khan KS, Daya S, Jadad A. The importance of quality of primary studies in producing unbiased systematic reviews. *Arch Intern Med* 1996; 156(6): 661-6.

Knipschild P, Kleijnen J, ter Riet G. Belief in the efficacy of alternative medicine among general practitioners in The Netherlands. *Soc Sc Med* 1990; 31(5): 625-6.

L'Abbe KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. *Ann Intern Med* 1987; 107(2): 224-33.

Laakmann G, Schule C, Baghai T, Kieser M. St. John's wort in mild to moderate depression: the relevance of hyperforin for the clinical efficacy. *Pharmacopsychiatry* 1998; 31 Suppl 1: 54-9.

Lambert MJ, Hatch DR, Kingston MD, Edwards BC. Zung, Beck, and Hamilton Rating Scales as measures of treatment outcome: a meta-analytic comparison. *J Consult Clin Psychol* 1986; 54(1): 54-9.

Le Bars P, Katz M, Berman N, Turan M, Freedman A, Schatzberg A. A placebo-controlled, double-blind, randomized trial of an extract of *Ginkgo biloba* for dementia. *JAMA* 1997; 278: 1327-32.

Lecrubier Y, Clerc G, Didi R, Kieser M. Efficacy of St. John's wort extract WS 5570 in major depression: a double-blind, placebo-controlled trial. *Am J Psychiatry* 2002; 159(8): 1361-6.

Lehrl S, Willemsen A, Papp R, WH. Ergebnisse von Messungen der kognitiven Leistungsfähigkeit bei Patienten unter der Therapie mit Johanniskraut. *Nervenheilkunde* 1993; 12: 281-4.

Levin JS, Glass TA, Kushi LK, Schuck JR, Steele LS, Jonas WB. Quantitative methods in research on complementary and alternative medicine: a methodological manifesto. *Med Care* 1997; 35: 1079-94.

Li Wan Po A. Dictionary of Evidence-based Medicine. Oxon: Radcliffe Medical Press Ltd, 1998.

Linde K, Mulrow CD. St John's wort for depression (Cochrane Review 2003). In: *The Cochrane Library* Oxford: Update Software.

Linde K, Ramirez G, Mulrow CD, Pauls A, Weidenhammer W, Melchart D. St John's wort for depression--an overview and meta-analysis of randomised clinical trials. *BMJ* 1996; 313(7052): 253-8.

MacLennan AH, Wilson DH, Taylor AW. Prevalence and cost of alternative medicine in Australia. *Lancet* 1996; 347(9001): 569-73.

Marshall RJ, Gee R, Israel M *et al.* The use of alternative therapies by Auckland general practitioners. *N Z Med J* 1990; 103(889): 213-5.

MIGHT. Greater Emphasis on Biotechnology under RM8 - A Boost for the Herbal Industry. Available at http://www.might.org.my/Activities/F_Greater.asp. (Accessed 27 May 2002).

Miller IW, Bishop S, Norman WH, Maddever H. The Modified Hamilton Rating Scale for Depression: reliability and validity. *Psychiatry Res* 1985; 14(2): 131-42.

Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control Clin Trials* 1995; 16(1): 62-73.

Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials. Current issues and future directions. *Int J Technol Assess Health Care* 1996; 12(2): 195-208.

Moller HJ. Rating depressed patients: observer- vs self-assessment. *European Psychiatry* 2000; 15(3): 160-72.

Moncrieff J. Are antidepressants overrated? A review of methodological problems in antidepressant trials. *J Nerv Ment Dis* 2001; 189(5): 288-95.

Montgomery SA. Can two week studies be used to establish efficacy? *Eur Neuropsychopharmacol* 1995; 5(3): 190-1.

Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry* 1979; 134:382-9.

Montgomery S. Clinically relevant effect sizes in depression. *Eur Neuropsychopharmacol* 1994; 4: 283-4.

Müller W. Current St. John's wort research from mode of action to clinical efficacy. *Pharmacological Research* 2003; 47(2): 101-9.

Muller WE, Rolli M, Schafer C, Hafner U. Effects of hypericum extract (LI 160) in biochemical models of antidepressant activity. *Pharmacopsychiatry* 1997; 30 Suppl 2: 102-7.

Muller WE, Rossol R. Effects of hypericum extract on the expression of serotonin receptors. *J Geriatr Psychiatry Neurol* 1994; Suppl 1: S63-4.

Murphy J, Heptinstall S, Doherty M, Mitchell J. Randomized double-blind, placebo-controlled trial of feverfew in migraine prevention. *Lancet* 1988; 2: 189-92.

Murray CL, Lopez AD. *The Global Burden of Disease*. Cambridge, Mass: Harvard University Press, 1996.

Murray EJ. Measurements issues in the evaluation of psychopharmacological therapy. In *The limits of biological treatments for psychological distress*. Hillsdale, NJ: Erlbaum, 1989: 39-68.

Newall CA, Anderson LA, Phillipson JD. Herbal medicines. A Guide for Health-care Professionals. 1st edition. London: Pharmaceutical Press, 1996.

Nicassio PM, Schuman C, Kim J, Cordova A, Weisman MH. Psychosocial factors associated with complementary treatment use in fibromyalgia. *J Rheumatol* 1997; 24(10): 2008-13.

NPCB. Report from Product Evaluation and safety Division. Annual Report 2000. [National Pharmaceutical Control Bureau (NPCB), Ministry of Health, Malaysia].

O'Connor BB, Calabrese C, Cardena E, Eisenberg D, Fincher J, Hufford D. Defining and describing complementary and alternative medicine. *Alternative Therapies* 1997; 3(2): 49-57.

Oppenheim AN. Designing attitude statements. In *Questionnaire Design, Interviewing and Attitude Measurement*. New edition., London: Printer Publishers Ltd, 1996.

Osiri M, Suarez-Almazor ME, Wells GA, Robinson V, Tugwell P. Number needed to treat (NNT): implication in rheumatology clinical practice. *Ann Rheum Dis* 2003; 62(4): 316-21.

Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med* 1992; 116(1): 78-84.

Perkin MR, Pearcy RM, Fraser JS. A comparison of the attitudes shown by general practitioners, hospital doctors and medical students towards alternative medicine. *Journal of the Royal Society of Medicine* 1994; 87(9): 523-5.

Persad E. Electroconvulsive therapy in depression. *Can J Psychiatry* 1990; 35(2): 175-82.

Peveler R, Carson A, Rodin G. ABC of psychological medicine: Depression in medical patients. *BMJ* 2002; 325(7356): 149-52.

Philipp M, Kohnen R, Hiller KO. Hypericum extract versus imipramine or placebo in patients with moderate depression: randomised multicentre study of treatment for eight weeks. *BMJ* 1999; 319(7224): 1534-8.

Physicians' Desk Reference (PDR). 53rd edition. Montvale, NJ: Medical Economics Company.

Quandt J, Schmidt U, Schenk N. Ambulante Behandlung leichter und mittelschwerer depressiver Verstimmungen. *Der Allgemeinarzt* 1993; 15(2): 97-102.

Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. *Applied Psychological Measurement* 1977; 1: 385-401.

Reh C, Laux P, Schenk N. Hypericum-Extrakt bei Depressionen - eine wirksame Alternative. *Therapiewoche* 1992; 42:1576-81.

Reilly DT. Young doctors' views on alternative medicine. *BMJ* 1983; 287(6388): 337-9.

Rey JM, Walter G. Hypericum perforatum (St John's wort) in depression: pest or blessing? *Med J Aust* 1998; 169(11-12): 583-6.

Rooney B, Fiocco G, Hughes P, Halter S. Provider Attitudes and Use of Alternative Medicine in a Midwestern Medical Practice in 2001. *Wisconsin Medical Journal* 2001; 100(7): 27-31.

Schempp CM, Pelz K, Wittmer A, Schopf E, Simon JC. Antibacterial activity of hyperforin from St John's wort, against multiresistant Staphylococcus aureus and gram-positive bacteria. *Lancet* 1999; 353(9170): 2129.

Schlich D, Braukmann F, Schenk N. Behandlung depressiver Zustände mit Hypericinium. *Psycho* 1987; 13: 440-7.

Schmidt U, Schenk N, Schwarz I, Vorberg G. Zur Therapie depressiver Verstimmungen. *Psycho* 1989; 15: 665-71.

Schmidt U, Sommer H. Johanniskraut-Extrakt zur ambulanten Therapie der Depression. *Fortschritte Medizin* 1993; 111: 339-42.

Schrader E. Equivalence of St John's wort extract (Ze 117) and fluoxetine: a randomized, controlled study in mild-moderate depression. *Int Clin Psychopharmacol* 2000; 15(2): 61-8.

Schrader E, Meier B, Brattström A. Hypericum treatment of mild-moderate depression in a placebo-controlled study. A prospective, double-blind, randomized, placebo-controlled, multicentre study. *Hum Psychopharmacol* 1998; 13: 163-9.

Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995; 273(5): 408-12.

Schulz KF, Grimes DA, Altman DG, Hayes RJ. Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. *BMJ* 1996; 312(7033): 742-4.

Shelton RC, Keller MB, Gelenberg A, Dunner DL, Hirschfeld R, Thase ME, Russell J, Lydiard RB, Crits-Cristoph P, Gallop R, Todd L, Hellerstein D, Goodnick P, Keitner G, Stahl SM, Halbreich U. Effectiveness of St John's wort in major depression: a randomized controlled trial. *JAMA* 2001; 285(15): 1978-86.

Sikand A, Laken M. Pediatricians' experience with and attitudes toward complementary/alternative medicine. *Arch Pediatr Adolesc Med* 1998; 152(11): 1059-64.

Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *J Clin Epidemiol* 1994; 47(8): 881-9.

Snapinn SM. Noninferiority trials. *Curr Control Trials Cardiovasc Med* 2000; 1(1): 19-21.

Snow V, Lascher S, Mottur-Pilson C. Pharmacologic treatment of acute major depression and dysthymia. American College of Physicians-American Society of Internal Medicine. *Ann Intern Med* 2000; 132(9): 738-42.

Sommer H, Harrer G. Placebo-controlled double-blind study examining the effectiveness of an hypericum preparation in 105 mildly depressed patients. *J Geriatr Psychiatry Neurol* 1994; 7 Suppl 1: S9-11.

Song F. Exploring Heterogeneity in Meta-Analysis: Is the L'Abbe Plot Useful? *J Clin Epidemiol* 1999; 52(8): 725-30.

Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol* 2001; 54(10): 1046-55.

Sterne JA, Egger M, Smith GD. Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis. *BMJ* 2001; 323(7304): 101-5.

Stewart LA, Parmar MK. Bias in the analysis and reporting of randomized controlled trials. *Int J Technol Assess Health Care* 1996; 12(2): 264-75.

Stewart LA, Parmar MK. Meta-analysis of the literature or of individual patient data: is there a difference? *Lancet* 1993; 341(8842): 418-22.

Thase ME, Greenhouse JB, Frank E *et al.* Treatment of major depression with psychotherapy or psychotherapy-pharmacotherapy combinations. *Arch Gen Psychiatry* 1997; 54(11): 1009-15.

Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994; 309(6965): 1351-5.

Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002; 21(11): 1559-73.

USP Press Release. USP announces criteria for levels of evidence policies for botanical articles. Available: http://www.usp.org/aboutusp/releases/2000/pr-2000-23.htm (Accessed: 2000, October 15).

John's wort or sertraline ? Randomized controlled trial in primary care. *Can Fam Physician* 2002; 48: 905-12.

Verhagen AP, de Vet HC, de Bie RA, Boers M, van den Brandt PA. The art of quality assessment of RCTs included in systematic reviews. *J Clin Epidemiol 2001*; 54(7): 651-4.

Verhoef MJ, Sutherland LR. Alternative medicine and general practitioners. Opinions and behaviour. *Can Fam Physician* 1995a; 41: 1005-11.

Verhoef MJ, Sutherland LR. General practitioners' assessment of and interest in alternative medicine in Canada. *Soc Sci Med* 1995b; 41(4): 511-5.

Villar J, Piaggio G, Carroli G, Donner A. Factors affecting the comparability of meta-analyses and largest trials results in perinatology. *J Clin Epidemiol* 1997; 50(9): 997-1002.

Vorbach EU, Arnoldt KH, Hubner WD. Efficacy and tolerability of St. John's wort extract LI 160 versus imipramine in patients with severe depressive episodes according to ICD-10. *Pharmacopsychiatry* 1997; 30 Suppl 2: 81-5.

Vorbach EU, Hubner WD, Arnoldt KH. Effectiveness and tolerance of the hypericum extract LI 160 in comparison with imipramine: randomized double-blind study with 135 outpatients. *J Geriatr Psychiatry Neurol* 1994; 7 Suppl 1: S19-S23.

Weissman MM, Klerman GL. Sex differences and the epidemiology of depression. *Arc Gen Psychiatry* 1977; 34(1): 98-111.

Weissman MMPhD, Bland RC, Canino GJ, Faravelli C, Greenwald S, Joyce PR, Karam EG, Lee CK, Lellouch J, Lepine JP, Newman SC, Rubio-Stipec M, Wells JE, Wickramaratne PJ, Wittchen H, yeh EK. Cross-National Epidemiology of Major Depression and Bipolar Disorder. *JAMA* 1996; 276(4): 293-9.

Wetzel MS, Eisenberg DM, Kaptchuk TJ. Courses involving complementary and alternative medicine at US medical schools. *JAMA* 1998; 280(9): 784-7.

Wharton R, Lewith G. Complementary medicine and the general practitioner. *BMJ* 1986; 292: 1498-500.

Wheatley D. LI 160, an extract of St. John's wort, versus amitriptyline in mildly to moderately depressed outpatients--a controlled 6-week clinical trial. *Pharmacopsychiatry* 1997; 30 Suppl 2: 77-80.

White AR, Resch KL, Ernst E. Complementary medicine: use and attitudes among GPs. *Family Practice* 1997; 14(4): 302-6.

Williams JB. A structured interview guide for the Hamilton Depression Rating Scale. *Arch Gen Psychiatry* 1988; 45(8): 742-7.

Williams JW, Mulrow CD, Chiquette E, Noel PH, Aguilar C, Cornell J. A systematic review of newer pharmacotherapies for depression in adults: evidence report summary. *Ann Intern Med* 2000; 132(9): 743-56.

Witte B, Harrer G, Kaplan T, Podzuweit H, Schmidt U. Behandlung depressiver Verstimmungen mit einem hochkonzentrierten Hypericumpräparat - eine multizentrische plazebokontrollierte Doppelblindstudie. *Fortschr Med* 1995; 113: 404-8.

Woelk H. Comparison of St John's wort and imipramine for treating depression: randomised controlled trial. *BMJ* 2000; 321(7260): 536-9.

Wong AH, Smith M, Boon HS. Herbal remedies in psychiatric practice. *Arch Gen Psychiatry* 1998; 55(11): 1033-44.

World Health Organisation. Definitions and conceptualisations of traditional medicine. Planning for cost- effective traditional health services in the new century - a discussion paper 1996; Available: http://www.who.or.jp/tm/introduction/bkg/3_definition.html (Accessed 6 March 2002).

World Health Organisation. The International Statistical Classification of Diseases and Related Health Problems. Tenth Revision (Volume 1), Geneva: World Health Organisation, 1992.

Zung WW. A self-rating depression scale. *Arch Gen Psychiatry* 1965; 12: 63-70.