Institute of Health Policy and Management

Erasmus University Rotterdam

Master Thesis

MSc Health Economics, Policy and Law

Investigating the quality of the evidence provided by empirical studies on women with breast cancer treated with mistletoe using GRADE (Grading of Recommendation, Assessment, Development, and Evaluation)

Student: Edeltraud J. Lenk

Student number: 362774

Supervisor: Hans Severens PhD

Co- readers: K. Redekop PhD and Maiwenn Al PhD

Expertise consultant: J. van den Bosch PhD, Pallas Health Research and Consultancy

Rotterdam, July 2013

Acknowledgements

This thesis was my final step towards graduation in the Master's programme in Health Economics, Policy and Law at IBMG, Erasmus University, Rotterdam.

Coming from Brazil, a country with still a long way to go regarding social justice, and also health inequalities, as a general practitioner I wanted to study a topic that could help dealing with these issues. Having also a background in Complementary and Alternative Medicine and believing that it can be one alternative to handle these and other conventional medicine problems, the choice of this topic had also the intention to improve my knowledge on how to further develop this area.

I would like to manifest my gratitude to my supervisor, Hans Severens, first for the willingness to supervise this topic, for all the invested time and patience in guiding my work, and helping me to gain experience, as well as a deeper and more critical view of the world of scientific reviews while writing my own.

I am also very grateful to Judith van den Bosch, Maiwenn Al and Ken Redekop, for all the support and the so many valuable comments, for the opportunity to learn a lot from the expertise of each one of them.

My gratitude reaches also Davina Marques, for her good sense of humour and the invaluable English proofreading.

I also want to thank my master colleagues, for making this important experience of an international master programme even more enriching.

Last, but not least, I deeply thank my husband, family and friends, for their love, patience and support during this time of studies and thesis writing.

Thank you all for the wonderful opportunity of learning a lot together!

Edeltraud Lenk

Rotterdam, July 22nd, 2013.

Contents

Acknowledgements ... 2
Contents ... 3
Abstract ... 4
List of abbreviations ... 5
Tables and Figures ... 6
Background ... 7
  Complementary and Alternative Medicine ... 7
  Anthroposophic medicine and research ... 7
  Mistletoe treatment ... 8
  Breast cancer ... 8
  Increasing interest and effort in research on CAM ... 9
  Importance of researching CAM ... 9
  Significance of showing strength of evidence of potential alternative intervention for important disease ... 10
    CAM researchers ... 10
    Health decision- and policy makers ... 10
    Practitioners and Patients ... 10
Objectives ... 10
Methods ... 10
  PICO ... 10
  Search ... 11
    Search strategy ... 11
    Selection ... 11
  GRADE system ... 12
  Applying GRADE ... 14
    Rating down ... 14
    Rating up ... 15
    Evidence Profile table and Summary of Findings table ... 15
  Data extraction (standardized excel sheets with GRADE criteria) ... 15
Results ... 16
  Studies ... 16
  Outcomes ... 20
  Completed standardized data extraction tables ... 21
  Evidence profile table ... 25
Discussion ... 27
  Study search: availability, German language, limitations ... 27
  Remarks when specifically assessing quality – experience of using GRADE ... 27
    Quality criteria ... 28
    Outcomes ... 29
    Difference in evidence quality between outcomes ... 31
    GRADE and complex interventions ... 31
Conclusion ... 33
Appendices ... 34
Reference list ... 51

Abstract

Background: With health costs rising and economic crisis affecting many countries, solutions that offer advantages for patients and society are most wanted, where Complementary and Alternative Medicine (CAM) appears to be calling the attention of politicians and governments. Among unconventional cancer therapies, mistletoe is one of the most frequently used, especially in Central Europe. Many reviews have addressed its safety, and effectiveness, with controversial results, mainly because conventional reviews have mostly included evidence coming from randomized controlled studies (RCTs). This review aims to use the GRADE (Grading of Recommendation, Assessment, Development, and Evaluation) approach to assess the quality of the evidence on mistletoe and breast cancer and to evaluate whether the GRADE method is more suitable for appraising evidence coming from CAM studies because of the innovations it proposes.

Methods: The databases PubMed/Medline, the Cochrane Library, SciVerse Scopus, Embase and CAMbase were searched for studies on mistletoe and breast cancer. The selected studies were grouped under six outcomes: survival, tumour progression, quality of life, immunological response, neutropenia, and safety. Quality was assessed regarding study limitations, imprecision, inconsistency, indirectness and publication bias, following GRADE guidelines, using standardized extraction sheets.

Results: Eleven studies (4 RCTs and 7 observational studies) met the inclusion/exclusion criteria and were included in the analysis. The complete assessment of initial and final grading of the quality of the evidence, the descriptions of the assessment and calculations of absolute and relative effects were only possible for safety, due to lack of data for the other outcomes. The quality of the evidence for safety, coming from 2 RCTs, was assessed as good and evidence from observational studies on survival and tumour response was graded as very low. RCTs on quality of life and immunological response were not assessable due to incompatible outcome measurements, which also happened with the studies on neutropenia.

Conclusion: The GRADE method proved to be an organized and transparent way of assessing quality of evidence. It treats many aspects differently from previous methods, for instance, considering evidence from observational studies and assessing quality per outcome. This makes it a good option for assessing evidence on CAM, often consisting of more observational studies than RCTs. The quality of the evidence on mistletoe shown by the GRADE approach was at least in line with evidence from other reviews. With the improvement of the quality of the observational studies on mistletoe, assessing this intervention with GRADE in the future might upgrade the quality of its evidence, providing better evidence for its recommendation.

List of abbreviations

BMJ	British Medical Journal
CAM	Complementary and Alternative Medicine
CI	Confidence Interval
CONSORT	Consolidated Standards of Reporting Trials
CTCAE	Common Terminology Criteria for Adverse Events
EBM	Evidence-Based Medicine
ECHAMP	European Coalition on Homeopathic and Anthroposophic Medicinal Products
EPHA	European Public Health Alliance
EU	European Union
GRADE	Grading of Recommendation, Assessment, Development, and Evaluation
HR	Hazard Ratio
HTA	Health Technology Assessment
ITT	Intention-to-Treat
MedDRA	Medical Dictionary for Regulatory Activities
NHIS	National Health Interview Survey (USA)
NHS	National Health Service (UK)
NICE	National Institute for Clinical Excellence
OIS	Optimal Information Size
OR	Odds Ratio
QoL	Quality of Life
RCT	Randomised Controlled Trial
RR	Risk Ratio
US	United States
WHO	World Health Organization

Tables and Figures

Figure 1.	Publications search flow chart	12
Table 1.	Summary of GRADE approach for downgrading and upgrading evidence	13
Table 2.	Summarized description of characteristics of selected studies	17
Table 3.	Studies grouped under each of the chosen outcomes	20
Table 4.	Standardized quality assessment tables showing GRADE criteria, used for judging each group of studies per outcome and here demonstrating the assessment of the evidence on safety	22
Table 5.	Final GRADE evidence profile table on adjuvant mistletoe treatment for breast cancer	26

Background

Complementary and Alternative Medicine

Even though there is no universally accepted definition of Complementary and Alternative Medicine (CAM), it can be described as 'a group of diverse medical and health care systems, practices, and products that are not generally considered to be part of conventional medicine'. The boundaries between CAM and conventional medicine are not absolute, and specific CAM practices may, over time, become widely accepted.[1]

Patients' use of CAM may differ, according to their country/culture, age, disease, among other factors.

A systematic review appointed an increase in CAM use and acceptance among the general public and medical personnel in 10 European countries from 1990 through 2006.[2] The CAMbrella project studied 18 of the 39 European member states and associated countries and gathered substantial research-based data revealing that the prevalence of CAM use may range from 0.3% to 86%, due to very few rigorous prevalence studies performed based on nationally representative samples (the vast majority of the studies are small and qualitatively poor).

The CAMbrella research has exposed EU citizens' demand for access to increased and diverse CAM provision, has shown that they face significant barriers in the access to CAM, that they wished for more support and information about CAM from conventional medical professionals, that they want access to trustworthy and reliable information to support informed decisions, and that they require transparent regulation of CAM training and practice.[3]

In the USA, the 2007 National Health Interview Survey (NHIS) found out that 83 million adults spent $33.9 billion out-of-pocket on CAM in a previous period of 12 months, which corresponded to 11.2% of total out-of-pocket expenditures on health care, and for approximately 1.5% of total health care expenditures.[4]

In England, this number is estimated to be GB£450 million per year (only about 10% covered by the NHS), while NHS expenditures for the same period for Family Health Services (prescribing costs not included) were £3846 million.[5]

In Australia, the numbers are as high as AU$4.13 billion, accountable for almost half the expenditures on nonsubsidized health care products.[6]

The use of CAM by cancer patients also depends on their context, of the cancer type, on the availability of the therapies, showing ranges from less than 10% to more than 80%,[7,8,9,10,11,12]

Anthroposophic medicine and research

Anthroposophic Medicine (AM) is an integrative diagnosis and therapy concept, an example of the integration of a holistic with a conventional approach. It combines mainstream scientific medicine with anthroposophy, a philosophy developed in the 20th century by Rudolf Steiner. AM considers a human being as a whole and aims to stimulate the self-healing forces of the body, restoring the balance of bodily functions and strengthening the immune system, rather than only or primarily relieving the symptoms of a disease. It relies not only on medications for treatment, but also on specific anthroposophic therapies, such as artistic therapies, eurythmy therapy, rhythmical massage, among others.[13]

Conventional medical research normally starts with understanding the biochemical and physiological mechanisms of a disease, and then developing a new chemical substance. In sequence, *in vitro* and animal testing is performed. Depending on the effects and risks measured by then, the four phases of clinical research in humans are conducted and, after licensing, the substance is finally integrated in clinical practice, often as a

"one-size-fits-all therapeutic prescription". As most CAM modalities, AM faces the inverse situation of having an already widespread clinical use of a treatment and then having to undergo safety, comparative effectiveness and component efficacy research - which often do not take the philosophy, processes and assumptions of the therapy into consideration - in order to understand the mechanisms through which treatments exercise their influence.[14,15]

In 2005, a Health Technology Assessment (HTA) report analysed efficacy, effectiveness, safety, utilization and costs of AM and was published as a book in 2006. [16] This report was updated in 2011, describing that efficacy and effectiveness were investigated by assessing 265 clinical studies (38 RCTs, 36 prospective and 49 retrospective non-randomized controlled trials, also 90 prospective and 52 retrospective trials without control groups) that included a wide range of anthroposophic treatments and various diseases. Most studies showed positive results for AM, and although methodological quality differed considerably, trials with better quality still showed a positive result and high external validity. [17]

Mistletoe treatment

The anthroposophic treatment of cancer patients with mistletoe's (*Viscum album)* extracts follows the AM principles previously described. These complex herbal preparations are prescribed for cancer in an attempt to restore the processes involved in the development of the disease, such as normal growth and apoptosis cycle of human cells, as well as the immunologic failure in recognizing and addressing it. This is accomplished by stimulating immunocompetent cells, and protecting the DNA of mononuclear cells.[18,19]

For this reason it is prescribed for almost all types of cancer, at various phases, and their common symptoms (as emaciation, fatigue and others), as well as for benign tumours, and other diseases related to immunological unbalance. The patient's individuality, involving biographic, emotional, dietetic, genetic and environmental elements is never ignored in the treatment, so dosage, frequency, and other aspects of the treatment differ due to individualized characteristics and response of each patient.[20,21]

Mistletoe is currently the best researched anthroposophic treatment. The first indications of mistletoe for cancer treatment in AM began in the 1920's. Since then, various research projects have been involved in investigating its effects on tumours, on the course of disease, its biological and pharmacological properties.

Oncologic research in mistletoe increased in the 1980's and the results of research of many international scientific institutions have been obtainable since then. [19,22] Among the various isolated pharmacologically active compounds, mistletoe lectins (ML I, II and III) and viscotoxins are responsible for mistletoe's cytotoxic effects. They are also the most exhaustively studied compounds. *V. album* chitin-binding agglutinin (VisalbCBA) is another pharmacologically important compound found in mistletoe, as well as flavonoids, triterpene acids, oligo and polysaccharides. Other acknowledged mistletoe effects are immune-modulating activity, DNA-stabilizing properties, improvements in quality of life (QoL), increased survival rate and lowered incidence of disease and conventional treatment adverse effects. [18,19,22,23,24,25]

Breast cancer

Breast cancer is responsible for a significant amount of morbidity and mortality in women.

The World Health Organization (WHO) figured the deaths by breast cancer to be 458,000 in 2008 and the diagnostics in 2010 to be around 1.5 million. [26] In the US, the estimation for breast cancer diagnostics and deaths in the year 2012 is 226,870 and 39,510, respectively. [27] Europe's incidence of breast cancer was estimated to be 450,000 and 139,000 deaths in 2008. [28]

High costs are associated with breast cancer treatment, with estimates of around US$13.9 billion being spent on breast cancer treatments every year in the US. [29]

A significant amount of patients diagnosed with breast cancer – 40 to 80% – use CAM in addition to the conventional prescribed therapy, mainly for support in controlling adverse effects of the conventional treatments, preventing or minimizing immunosuppression, for better quality of life and also to prolong life. [23] Lower costs of in-patient treatment, as well as lower economic loss in productivity are also attributed to the use of mistletoe preparations in the breast cancer aftercare. [30,31]

Increasing interest and effort in research on CAM

The European Union has become increasingly interested in CAM, as the CAMbrella project within the 7th EU Framework Programme has shown [32,33], as well as the several events that have been taking place in the European Parliament. On October 9th 2012, the evidence-based relevance of CAM for the future development of the public health agenda and health care delivery in the EU was discussed. On June 27th 2013, the topic was the role of CAM, its provision and integration to health care systems and suitability for the EU's current "Investing in Health" policy. In September 2013 the debate will be on CAM's contribution to the improvement of health outcomes. Institutions such as the European Public Health Alliance (EPHA), ECHAMP, Cochrane Collaboration's CAM field, and several European Parliament Interest Groups (such as MEPs Against Cancer and MEPs for CAM) have maintained the debate and the inclusion of CAM in research as well as in public health discussions. [34,35,36]

Importance of researching CAM

Some of the reasons that illustrate and legitimate the growing importance of research in the CAM field include:

- The ethical need to assess the safety and effectiveness of any health care interventions that can contribute to the health care market (most important in times of economic crisis), but of CAM interventions specifically, which lack research on these topics;
- The significant scientific interest to ascertain whether and how specific complementary and alternative therapies work;
- The fact that evidence-based integrative medicine can only be developed with scientific evidence on CAM;
- The understanding that politicians, policy-makers, stakeholders need reliable information to be able to decide upon licensing, reimbursing, recommending, discouraging or banning therapies;
- The consideration of research as 'the systematic gathering of data, information and facts for the advancement of scientific knowledge' leads us to the idea that research can and should result in improvement of clinical care. Questioning theories and processes – established or not – and developing new ideas can help improving health care;
- The patient's demand for more and more diverse CAM provision, for CAM availability in the normal healthcare, for CAM provision by therapists with specific training;
- The importance of providing further arguments for the discussions about assuring patient's right of choice, as well as equality, CAM legalization and regulation.

But, in order to be helpful, research in CAM therapies has to be highly qualified, and should primarily reflect the interests of patients and society. [3,14,32,35]

Significance of showing strength of evidence of potential alternative intervention for important disease

CAM researchers

The results will be important to further develop CAM research, showing where it is most commonly downgraded or upgraded for quality issues, and consequently indicating how to improve future research (if the purpose is to fit the current scientific paradigm).

Health decision- and policy makers

The undeniable current demand for health systems to improve quality and efficiency - above all in times of financial crises – makes CAM a possible option for needed solutions, also considering the importance CAM has for patients/citizens. Therefore, reliable and transparent ways of showing the quality of the evidence that could lead to future recommendations and guidelines including CAM need to be explored.

Showing the quality of the evidence of these studies by outcome is also not the common approach in reviews, but, given CAM's potential contributions to health systems, this outcome-centred approach could be more easily translated into recommendations.

Practitioners and Patients

Health practitioners and patients will have more data to discuss treatment options and make safe and proper health care choices for specific clinical circumstances.

Objectives

This thesis proposes to investigate the quality of the evidence provided by comparative empirical studies on women with breast cancer treated with mistletoe by using the GRADE approach. It also aims to examine if and how the GRADE approach is more suitable for appraising evidence coming from CAM studies compared to previous methods, for the innovations it proposes.

Methods

PICO

The **PICO** method was applied to formulate the research question and to determine the research scope. The **P**opulations consists of breast cancer patients, all stages included. The **I**ntervention is mistletoe extract administered in addition to conventional treatment, mainly subcutaneously, 2-3 times per week, with variable dosage. The **C**omparison was conventional treatment (surgery, radiation, hormonal therapy and/or chemotherapy) alone. The **O**utcomes were safety (adverse effects), survival, quality of life, tumour progression/response, positive immunologic response and neutropenia. On the appropriateness of patients, interventions, and outcomes to include, and on the combination of results across studies, GRADE suggests that reviewers should start by pooling widely, and later testing if the assumption of similar effects across studies holds.[37]

Search

Search strategy

Between July and August 2012 the following databases were searched: PubMed/Medline, the Cochrane Library, SciVerse Scopus, Embase and CAMbase. A list of publications about anthroposophic medicine from 2005 to 2011, held by the Anthroposophic Medical Section was reviewed, as well as reference lists of relevant review articles on the topic. The separate search terms were “breast” and “cancer” and “mistletoe” or “viscum” or “iscador” or “helixor” or “abnoba” or ”isorel” or “visorel” or “eurixor” or “lektinol” or “iscucin” or “plenosol”.

Selection

For inclusion in the analysis, the following selection criteria were used: any controlled study design, study population with any type of breast cancer, intervention group treated with any mistletoe preparation, clinically relevant patient outcome, completion and publication of study.

The exclusion criteria were: reviews, articles written in other language than English, articles that included also other types of cancer or other types of therapies, only abstracts, double publication (exception for presentation of further data), and *in vitro*/animal experiments.

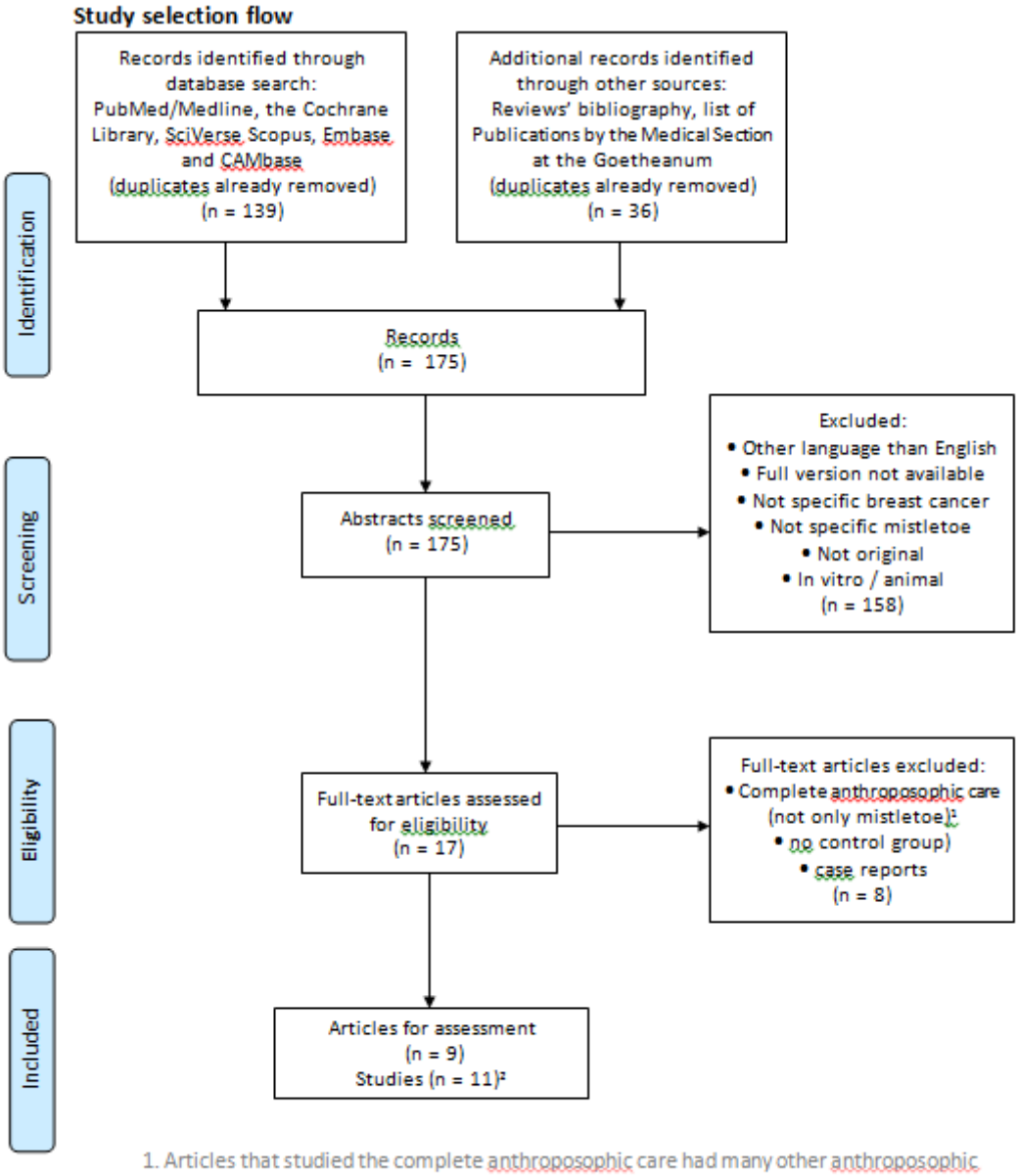

Figure 1. Publications search flow chart

GRADE system

The Grading of Recommendation, Assessment, Development, and Evaluation (GRADE) Working Group comprises researchers, health professionals, and guideline developers since 2000, in a worldwide and continuous effort to elaborate an optimal system for rating quality of evidence and specifying the strength of recommendation for clinical practice guidelines.[38]

More than merely a rating system, GRADE provides a comprehensive, structured and transparent process for performing quality assessment and developing clinical and practical recommendations and its use is applicable and helpful regardless of the quality of the evidence: whether high or low. It considers the current conventional evidence-based hierarchy pyramid, but also discusses the weight the quality of the evidence has

for the assessed intervention when reviewing or making recommendations about it. Moreover, it separates the evidence per outcome, which enables also a different assessment of quality, since quality can clearly differ within outcomes in the same study. [37,39]

Currently, 74 institutions have endorsed or are using GRADE, such as the World Health Organization (WHO), the Cochrane Collaboration, the National Institute for Clinical Excellence (NICE), the Dutch Institute for Healthcare Improvement – CBO, the British Medical Journal (BMJ), the Robert Koch Institute (Germany), the Norwegian Knowledge Centre for the Health Services, the Swedish National Board of Health and Welfare, the Spanish Society for Family and Community Medicine, the American College of Physicians, among others. This increasing number of respected organizations adhering to the GRADE methodology has also supported the choice for this assessment system in this dissertation.[40]

The current medical scientific paradigm follows David Sackett's EBM (Evidence-Based Medicine) definition: "conscious, explicit and sensate use of the best evidence available in decision making about patient care, added to the physician's experience and the patient's preferences".[41] For various reasons, the best available evidence for many interventions is of low or very low quality, and GRADE also enables this level of evidence to be considered when making recommendations.

Another GRADE dimension that could specifically suit CAM and many conventional practices that have difficulties performing RCTs to show their effects (i.e. psychology, physiotherapy, surgery) is the possibility to upgrade the quality of the evidence provided by methodologically rigorous observational studies from low to moderate or even high.[42]

Table 1. Summary of GRADE approach for downgrading and upgrading evidence

Source of body of evidence	**Initial rating of quality of a body of evidence**	**Factors that may decrease quality**	**Factors that may increase quality**	**Final rating**
Randomized trials	High	1. Study limitations (Risk of bias)	1. Large magnitude of effect	High - further research is very unlikely to change the confidence in the estimate of effect
Observational studies	Low	2. Inconsistency	2. Dose-response gradient	Moderate - further research is likely to have an important impact on the confidence in the estimate of effect and may change the estimate
		3. Indirectness	3. All plausible residual confounders or biases increase the confidence in the estimated effect	
		4. Imprecision		Low - further research is very likely to have an important impact on the confidence in the estimate of effect and is likely to change the estimate
		5. Publication bias		Very low – any estimate of effect is very uncertain

(adapted from Brozek JL, Bousquet J, Baena-Cagnani CE, Bonini S, Canonica GW et al.) [43]

Applying GRADE

The GRADE guidelines are applied somewhat differently for systematic reviews than for guidelines, starting from the definition of "quality". For systematic reviews, quality is referred to as the confidence in the estimates of effect. For guidelines, it is the extent to which the confidence in the effect estimate is adequate to endorse a specific decision. This assessment used the systematic review approach in all the steps.[44]

The first suggestion of the GRADE guidelines is to explicitly define the question that the assessment is addressing following PICO, in this case: "Should mistletoe be used in addition to conventional treatment for breast cancer?" Next, the important outcomes are specified, which should include harms, and all outcomes important to patients.[37]

In this assessment, the outcomes were defined after the study selection, among the common outcomes of the selected articles, in order to enable the use of GRADE for several different outcomes. After grouping the studies per outcome, each group is assessed for its quality of evidence, following the criteria described below. In the GRADE approach, randomized trials start as high and observational studies as low quality evidence. [39,44]

Rating down

Study limitations (risk of bias) are assessed differently depending on the study design.

For RCT assessment, the following criteria are used: allocation concealment, blinding, complete accounting of patients and outcome events, selective outcome of reporting bias, stopping early for benefit, use of unvalidated outcome measures, carryover effects in crossover trial, recruitment bias in cluster-randomized trials, and baseline values balance.

For observational studies, these criteria are: failure to develop and apply appropriate eligibility criteria (inclusion of control population); under- or overmatching in case-control studies; selection of exposed and unexposed in cohort studies from different populations; flawed measurement of both exposure and outcome; differences in measurement of exposure; differential surveillance for outcome in exposed and unexposed in cohort studies; failure to adequately control confounding; failure of accurate measurement of all known important prognostic factors; failure to match for prognostic factors and/or lack of adjustment in statistical analysis, and incomplete follow-up.

Even though the GRADE guidelines suggest that authors should consider including only studies with a lower risk of bias, [45] this assessment also included studies with higher risk of bias, for the sake of using the approach for exercising judgment on downgrading and in order to understand how far it should be done.

Within a RCT, the risk of bias might be classified as low for all key criteria, crucial limitation for one or some limitations for multiple criteria, and crucial limitation for more criteria. Across studies, it is possible to find most of the information coming from studies at low, moderate and high risk of bias. This will determine whether the limitations will be considered as not serious (do not downgrade), serious (rate down one level) or very serious (rate down two levels)**.**[45]

Imprecision is considered as 'not serious' if the Optimal Information Size (OIS) criterion is met and if the 95% Confidence Interval (CI) excludes no effect (if CI around RR excludes 1.0), demanding no rating down. Otherwise, imprecision can be considered serious or very serious and therefore rate down one or two levels.[46]

The criteria for assessing **inconsistency** in results are similarity of point estimates, CI overlap extent, statistical test for heterogeneity (tests the null hypothesis that all included studies have the same inherent magnitude of effect), and I^2 (defines the percentage of the variation in point estimates imputable to among-study

differences). Description terms used are also: not serious, serious and very serious and downgrading performed accordingly.[47]

Indirectness is judged by differences in population and intervention (applicability), differences in outcome measures (surrogate outcomes), indirect comparisons because of biased head-to-head comparisons (e.g. industry), and mechanism of action (applicability, believing analyses, surrogate outcomes).[48]

The extent the chosen studies show early positive studies, small in size, preliminary and pilot studies; non-English publications, double counting, industry sponsorship (or likely to be) or conflict of interest, gives the likelihood of **publication bias**. The description terms suggested by GRADE are "undetected" or "strongly suspected", and GRADE suggests rating down a maximum of one level (due to difficulty in assessing the likelihood of publication bias). Due to the same difficulty in assessing the likelihood, the description terms used in this review will be: possible or unlikely.[49]

Rating up

Three criteria are included in the GRADE method for rating up quality of evidence, which are specially applicable to observational studies: a large (associated RR from 2 to 5) or a very large (associated RR > 5) magnitude of effect with no plausible confounders and no relevant problems with risk of bias of precision; a dose-response gradient and/or the conclusion that plausible residual confounding or biases would reduce the proven effect or suggest effect when results showed none.[42]

Evidence Profile table and Summary of Findings table

For systematic reviews and HTA limited to evidence reports, the endpoint of the GRADE process is a summary of the evidence. It is presented as an evidence profile table showing the number of studies and study design, the judgment for each previously-mentioned criterion (with the correspondent justification as a footnote), and a summary of findings with the best estimates of magnitude of relative and absolute effects. For guideline developers, this summary is a crucial step on the way to a recommendation.[39]

For the measures of relative effect, RR is preferred over OR, because of being more intuitively understandable. For measures of absolute effect, the baseline risk would come ideally from well-designed observational studies. In case they are not available, it would be calculated from the median risk (rather than weighted average) from control groups in the included studies. Absolute effects should be presented as natural frequencies (events per 10,000 patients, if they are more frequent then per 1,000 or 100 patients), in order to facilitate decision making. The presentation should be consistent across all outcomes in one table.[50]

GRADE then requires the rating of confidence in estimates of effect (quality of evidence) in high, moderate, low or very low, for each outcome. Each rating is also disclosed with the respective description of the rationale for the final decision as a footnote.[39]

Guideline developers will still go a step further, and, after choosing which outcomes are critical for assessing this intervention, they will make an overall rating of confidence in effect estimates across all outcomes. The recommendation will be based on this rating. [51] This review did not perform this last step, for it was not meant for recommendations. Refer to the Results section for the evidence profile table.

Data extraction (standardized excel sheets with GRADE criteria)

Data extraction was made using standardized Excel sheets including the previously listed criteria, listing the studies grouped for each of the chosen outcomes and the description of how the studies performed at each

criterion (refer to Results section for the complete tables on safety and to the Appendices for the missing tables).

Results

Studies

After applying the inclusion and exclusion criteria, 9 articles were included in the review. Two studies are described in each of the two articles by Grossarth-Maticek, leading to a final number of 11 studies. Of these, 4 are RCTs and 7 are observational studies, being 9 prospective, one retrospective and one retrolective ('characterized by the sampling of anonymous data from medical records in standardized case report forms and by a follow-up starting from the origin (i.e. diagnosis or primary surgery) in the past, with pre-specified endpoint(s) in the past, present or future.').

With exception of Büssing, that used an intravenous infusion, all studies used mistletoe in the subcutaneous form. The 3 most recent studies used the escalating dose approach, mostly 3 times a week; 5 used dosage and frequency according to patients' conditions and the opinion of the attending physician; 2 used fixed doses 2 times a week and one study used one single intravenous infusion.

The studies were mostly conducted in Europe (Germany, Switzerland, Austria, Russia, Serbia, Bulgaria), and only one in Asia (Korea). The study settings were mainly oncologic research centra, oncologic hospitals, and academic hospitals.

Except for Büssing's 3 day study, follow-up ranged from 7 weeks to more than 5 years. The outcomes included safety/adverse effects, survival, quality of life (QoL), efficacy, immunologic response, neutropenia, prevention of surgery-induced suppression of granulocyte function, psychosomatic self-regulation, use of antiemetic and analgesic drugs, number of inpatient days, Karnofsky scale. A summary of the characteristics of the selected studies can be seen in Table 2.

Table 2. Summarized description of characteristics of selected studies

Author	Study Design	Location	n		Breast cancer population	Intervention	Dose	Comparison	Follow-up	Outcomes[4]
			Test	Control						
Son 2010 [52]	RCT (not blinded)	Korea	10	10	UICC stage I or II[3] invasive ductal carcinoma, 33 to 63 years, estrogen receptor + or -	conventional surgery + chemotherapy + radiotherapy + Helixor (aqueous extract)	subcutaneous 3 x / week escalating dose from 1 to 100 mg (1mg = aprox 40-60 ng lectin)	conventional surgery + chemotherapy + radiotherapy - no mistletoe	7 weeks	immunologic response (IL-2, IL-4, IL-6, IL-10, TGF-β and IFN-γ)
Tröger 2009[53]	RCT (prospective randomized open label study)	Serbia	30	31	UICC stages I – III[3], older than 18 years, clinically well	conventional surgery+ chemotherapy without radio- or hormone therapy + Iscador M special (lacto-fermented aqueous extract)	subcutaneous 3 x / week escalating dose from 0.01 to 5 mg	conventional surgery+ chemotherapy without radio- or hormone therapy - no mistletoe	average 18 weeks	quality of life; neutropenia; safety
Beuth 2008[54]	Observational (retrospective comparative cohort)	Germany	167	514	UICC levels I-III[3], age 20-80	conventional surgery + chemotherapy + radiotherapy + hormone therapy + Helixor (aqueous extract)	subcutaneous 2-3 x/week escalating dose from 1 to 50 mg	conventional surgery + chemotherapy + radiotherapy + hormone therapy - no mistletoe	intervention 4.35 years; control 3.0 years (mean)	Safety; efficacy (relief of typical disease or conventional therapy-induced symptoms)
Grossarth-Maticek 2006 A[1] [55]	Observational (cohort) - prospective randomized matched-pair study	Germany	38	38	without any recurrence or metastases	conventional treatment as applicable: surgery, chemotherapy, radiotherapy and hormone therapy + Iscador (lacto-fermented aqueous extract)	subcutaneous 2-3 x/week, no info about doses	conventional treatment as applicable: surgery, chemotherapy, radiotherapy and hormone therapy - no mistletoe	at least 1 year (until 1998 or death)[5]	overall survival; psychosomatic self-regulation; tumour progression

	Observational (cohort) – prospective non-randomized matched-pair study		84	84	without any recurrence or metastases	conventional treatment as applicable: surgery, chemotherapy, radiotherapy and hormone therapy) + Iscador (lacto-fermented aqueous extract)	subcutaneous 2-3 x/week, no info about doses		at least 1 year (until 1998 or death)[5]	
Grossarth-Maticek 2006 B[2] [56]	Observational (cohort) – prospective randomized matched-pair study	Germany	17	17	only lymphatic metastases, pre and postmenopausal	Conventional (any combination of chemo/ hormone/ radio therapies, surgery) + Iscador (lacto-fermented aqueous extract)	subcutaneous 2-3 x/week, no info about doses	conventional (any combination of chemo/ hormone/ radio therapies, surgery) - no mistletoe	at least 3 weeks (until 1998 or death)[5]	overall survival; psychosomatic self-regulation
	Observational (cohort) – prospective non-randomized matched-pair study		180	180	42 pairs local recurrences and no metastases, 55 pairs only lymphatic metastases, 83 pairs distant metastases, pre and post-menopausal	conventional (any combination of chemo/ hormone/ radio therapies, surgery) + Iscador (lacto-fermented aqueous extract)	Subcutaneous 2-3 x/week, no info about doses		at least 3 weeks (until 1998 or death)[5]	
Semiglasov 2006 [57]	RCT (double blind)	Russia, Bulgaria and Ukraine	176	176	UICC stages I-II[3,] including carcinoma in situ, 18 - 55 years, pre and postmenopausal	chemotherapy + Lektinol (aqueous extract)	subcutaneous 2x/ week 15 ng mistletoe lectin	chemotherapy + placebo	24 to 32 weeks	quality of life; haematologic parameters; Karnofsky scale; consumption of antiemetic and analgesic drugs; number of inpatient days; adverse effects

Büssing 2005 [58]	Phase II Clinical trial (controlled, prospective, open label, non-randomized)	Germany	47	51	any stage with planned surgical intervention, 18 - 80 years	standard anaesthesia + Iscador M special, (lacto-fermented aqueous extract)	1mg intravenous (aprox 40 – 60 ng lectin)	standard anaesthesia - no mistletoe	3 days	prevention of surgery-induced suppression of granulocyte function
Semiglasov 2004 [59]	RCT (double blind)	Russia, Bulgaria and Ukraine	202 (65)[6]	70	operable breast cancer and eligible for adjuvant chemotherapy, UICC stages I-II[3], 18 - 55 years, pre and postmenopausal	chemotherapy + Lektinol (aqueous extract)	subcutaneous 2x/ week 10, 30 or 70ng mistletoe lectin/ml	chemotherapy + placebo	15 weeks	quality of life; haematology; consumption of antiemetic and analgesic drugs; number of inpatient days; safety; immunological parameters
Bock 2004 [60]	Observational (retrolective comparative cohort with parallel groups)	Germany and Switzerland	710	732	post-surgical, primary, non-metastatic	conventional treatment (as applicable, surgery, chemotherapy, radiotherapy and hormone therapy) + Iscador (lacto-fermented aqueous extract)	subcutaneous, dose and frequency according to the individual patient's health status and preference at the discretion of the treating physician	conventional treatment (as applicable, surgery, chemotherapy radiotherapy and hormone therapy) - no mistletoe	at least 3 years or until patient's death	efficacy; overall survival during study and follow-up); safety

1. Prospective Controlled Cohort Studies on Long-Term Therapy of Breast Cancer Patients with a mistletoe preparation (Iscador). Grossarth-Maticek R, Ziegler R, *Forschende Komplementärmedizin* 2006; 13:285-292.

2. Randomised and Non-randomised Prospective Controlled Cohort Studies in Matched-pair Design for the Long-Term Therapy of Breast Cancer Patients with a mistletoe preparation (Iscador): A Re-analysis. Grossarth-Maticek R, Ziegler R, *European Journal of Medical Research* 2006; 11:485-495.

3. According to the International Union against Cancer TNM staging system (Tumour/Nodes/Metastasis).

4. Primary and secondary outcomes.

5. No average/mean for follow-up duration available

6. Only the medium dose group was included in the Quality of Life assessment because the dose is in line with the other studies and because no statistical relevant results were shown in the other groups.

Outcomes

Among the outcomes described before, 6 were chosen that are patient important, and that had more than one study. For safety assessment 5 studies were available, 2 for tumour response, 3 for QoL, 2 for immunologic response and 3 for neutropenia. Table 3 shows which studies were grouped under each outcome and the correspondent outcome measurement.

Table 3. Studies grouped under each of the chosen outcomes

Outcome	Study	Outcome measures
Survival	Grossarth-Maticek 2006 A	time from first diagnosis to death, HR for overall mortality
	Bock 2004	HR for overall mortality
	Grossarth-Maticek 2006 B	time from first diagnosis to death, HR for overall mortality
Tumour progression/response	Grossarth-Maticek 2006 A[1]	time from first diagnosis to local recurrences, lymphatic metastases or distant ones
Quality of Life	Tröger 2009	EORTC-QLQ-C30 in the official Serbian translation, before each CAF cycle and three weeks after the 6th CAF cycle (= 7 visits)
	Semiglasov 2006	FACT-G scale as primary efficacy variable and GLQ-8 and Spitzer's uniscale secondary outcome variables
	Semiglasov 2004	GLQ-8 and Spitzer's uniscale + EORTC QLQ C30 (no results shown on this score – non-significant differences) prior to each CMF cycle and 2 and 3 weeks after the 4th CMF cycle
Safety	Beuth 2008	adverse effects: local reactions (erythema, pruritus) and systemic reactions (flu-like symptoms) attributed to test treatment
	Tröger 2009	assessment of adverse effects
	Semiglasov 2006	adverse events attributed to test treatment
	Semiglasov 2004	adverse events attributed to test treatment
	Bock 2004	adverse drug reactions attributed to test treatment
Positive immunologic response	Son 2010	IL-2, IL-4, IL-6, IL-10, TGF-β and IFN-γ
	Semiglasov 2004	NK-cell activity; count: NK-cells, CD3+HLA DR+, CD4+, CD8+, CD4+/CD8+ ratio, CD25+, MAC1+
Neutropenia	Tröger 2009	neutropenia defined as neutrophil count <1,000/µl in peripheral blood, one day before each CAF cycle and three weeks after the 6th CAF cycle
	Semiglasov 2006	minimum values in the course of treatment (parameters not specified); time of first occurrence of leucopenia/granulocytopenia
	Büssing 2005	surgical induced suppression of granulocyte function measured by oxidative burst of granulocytes prior to Viscum application, days 1 and 3, by flow cytometry, with E.coli or PMA stimulation (phorbol-12-myristat-13-acetate)

1. This publication includes 2 observational studies (randomized and non-randomized)

Completed standardized data extraction tables

Table 4 is a resume of the standardized quality criteria used in the assessment of each group of studies per outcome, according to the GRADE guidelines: study limitations (risk of bias) for both observational studies and RCTs, inconsistency, imprecision, indirectness and publication bias. The tables show the example of the assessment of the evidence on safety, since it was the only outcome possible to be assessed in all criteria.

The complete tables with the judgments, calculations and comments can be found in the Appendices.

Table 4. Standardized quality assessment tables showing GRADE criteria, used for judging each group of studies per outcome and here demonstrating the assessment of the evidence on safety

Observational												
Study	**Design**	**Limitations (risk of bias / internal validity)**										
		Failure to develop and apply appropriate eligibility criteria (inclusion of control population)	Under- or overmatching in case-control studies	Selection of exposed and unexposed in cohort studies from different populations	Flawed measurement of both exposure and outcome	Differences in measurement of exposure	Differential surveillance for outcome in exposed and unexposed in cohort studies	Failure to adequately control confounding	Failure of accurate measurement of all known important prognostic factors	Failure to match for prognostic factors and/or lack of adjustment in statistical analysis	Incomplete follow-up	Result
Bock 2004	Retrolective Comparative Cohort (parallel group design)	yes = no inclusion of control group in the adverse events assessment.	no	no	yes = no clear description of adverse events (only mentions common toxicity criteria and separates by local and systemic drug adverse reaction only), no separate presentation of the different kinds of adverse events	yes = adverse events not measured in control group	yes = no surveillance for adverse effetcs in control group	yes = confounding for mistletoe common adverse events like fever/chills not considered	no - tables 1 and 2 (pg 27) - quite complete, test group even showed more severe and advanced disease	no - multivariate adjusting and stratification	**intention-to-treat analysis** + 10.3% excluded for severe protocol violation (sensitivity analysis did not detect significant bias on the adjusted outcomes due to exclusion)	Very serious limitations
Beuth 2008	Retrospective Comparative Cohort	yes = no inclusion of control group in the adverse events assessment.	no	no	yes = no clear description of adverse events, no separate presentation of the different kinds of adverse events	yes = adverse events not measured in control group	yes = no surveillance for adverse effetcs in control group	yes = confounding for mistletoe common adverse events like fever/chills not considered	no - page 525, test group showed a tendency to more advanced disease (higher UICC stages) and more frequent hormone-therapy in the control group (advantages for control group)	no - logistic regression analysis	2.4% study group and 1.8% control group (died)	

RCT													
Study	**Design**	**Limitations (risk of bias / internal validity)**											
		Lack of allocation concealment	Lack of blinding	Incomplete accounting of patients and outcome events	Selective outcome reporting bias	Stopping early for benefit	Use of unvalidated outcome measures	Carryover effects in crossover trial	Recruitment bias in cluster-randomized trials	Baseline values	Unappropriate consideration of Intention-to-treat principle	Conclusion	Result
Tröger 2009	Prospective randomized open label clinical trial (pilot study)	Low risk. Probably done. Quote: "Allocation concealment was implemented by using sealed envelopes" (…)	High risk. Probably not done. Quote: " CRDT (Clinical Research Dr. Tröger) was responsible for planning, conduct, monitoring, and analysis of the study." Reason for not placebo-controlled: "typical reactions following injections of mistletoe extracts cannot be imitated by a pseudo-placebo".	Low risk. One patient control group withdrawn for heart disease (3%); less than 1% missing data at both groups for both assessed parameters	High risk. Planned outcomes reported, but methods section suggests that other adverse events then injection site reaction were measured, but no information on this issue (i.e. flu-like symptoms), no results presented for adverse events in control group.	Low risk.	Low risk (absent).	Low risk (absent).	Absent	Balanced (pg 38)	Absent (pg37: "statistical analysis was performed on the intention-to-treat population").	Crucial limitation for two criteria.	Excluded from analysis
Semiglazov 2006	RCT (double blind)	Low risk. Probably done. Quote: "Patients were allocated to the treatment groups on the basis of a computer-generated randomisation list."	Low risk. Double blind.	Low risk. 4 patients lost to placebo group = 2% (2 decision of the patient, 1 serious adverse effect, 2 patients moved). 1 patient lost to test group = 0,6% (decision of patient)	High risk. Planned outcomes reported, but adverse events and safety laboratory tests not specified at methods section.	Low risk.	Low risk (absent).	Low risk (absent).	Absent	Balanced between treatment groups with respect to demographics and medical history, not with respect to QoL scale scores (stronger restrtiction at test group) and histological classification (lower rate of invasive tumors inplacebo group), evaluation complemented by adjusted analyses of variance.	All patients included in the evaluation.	High risk of bias for one criterion	No serious limitations
Semiglasov 2004	RCT (double blind)	Low risk. Probably done. Quote: "Patients were allocated to the treatment groups on the basis of a computer-generated randomisation list."	Low risk. Double blind.	Low risk. 4% (11 patients lost: 4 to adverse effects, 4 decision of patient, 3 other reasons) ITT analysis.	High risk. Planned outcomes reported, but adverse events and safety laboratory tests not specified at methods section.	Low risk.	Low risk (absent).	Low risk (absent).	Absent	Balanced between treatment groups with respect to demographics and medical history, not with respect to QoL scale scores (not statistically significant)	Absent (ITT analysis performed).	High risk of bias for one criterion	

Study	Design	Inconsistency - In relative, not absolute, measures of effect (RR/HR/OR) - Rate down if:					Imprecision (around absolute, rather than		
		Point estimates vary widely across studies	Confidence intervals (Cis) show minimal or no overlap	Statistical test for heterogeneity shows a low p-value (<0.10 suggested by Cochrane for few studies instead of <0.05)	I^2 is large	**Result**	If the OIS (optimal information size) is not met, rate down for imprecision unless sample size very large.	95% Confidence Interval excludes no effect (CI around RR excludes 1.0)	**Result** If OIS met and 95% CI around effect excludes 1.0, no need to rate down for imprecision.
Semiglazov 2006	RCT (double blind)	no	Confidence intervals show reasonable overlap	p = 0.7052	I^2 is zero	**No serious inconsistency**	OIS met (calculated OIS = 56)	No effect excluded	**No serious imprecision**
Semiglasov 2004	RCT (double blind)								

Study	Design	Indirectness						Publication Bias					
		Differences in population (applicability)	Differences in interventions (applicability)	Differences in outcome measures (surrogate outcomes)	Indirect comparisons because of biased head-to-head comparisons (e.g. industry)	Mechanism of action	Result	Early positive studies, small in size, preliminary and pilot studies	Non-English speaking countries might submit negative studies to local journals, "gray literature"	Double counting	Industry sponsored (or likely to be) or conflict of interest	Assymetrical funnel plot	Result
Observational													
Bock 2004	Retrolective Comparative Cohort (parallel group design)	not the case	not the case	not the case	not the case	not the case	no serious indirectness	no	possible	not the case (revised for dupplicates)	all studies performed by research institutes (public, universities, Bock=independent research institute)	not possible	Possible Publication bias
Beuth 2008	Retrospective Comparative Cohort												
RCT													
Semiglazov 2006	RCT (double blind)	not the case	not the case	not the case	not the case	not the case	no serious indirectness	small in size, not early positive, Semiglasov 2004 pilot	possible	not the case (revised for dupplicates)	both received funding/medication provision from Madaus GmbH, but were performed by independent research centra	Not considered (too few studies)	Possible Publication bias
Semiglasov 2004	RCT (double blind)												

Evidence profile table

Table 5 shows the final evidence profile table on mistletoe therapy for breast cancer based on 2 RCTs and 2 observational studies, which is the end point of the GRADE process for systematic reviews.

GRADE encourages the summary of evidence from both RCT and observational studies in the evidence profile table when both provide important evidence with similar confidence in estimates. [50] This review expected to be able to experience providing evidence from both RCTs and observational studies and also to eventually experience rating up evidence from observational studies on safety, which is another GRADE feature. Unfortunately, the lack of data on adverse effects incidence in the control group in the 2 observational studies made it unfeasible to disclose information of both study designs. Consequently, it also resulted in the rating down of the quality of the evidence, excluding it from the assessment on safety. The same happened with the study by Tröger, which was also excluded from the RCT group for not disclosing any information on adverse events for the control group.

Therefore, safety shows the result of a complete assessment, coming from 2 RCTs, in this case not rated down and judged as high quality.

The safety results on the summary of findings table make it seem discouraging to recommend mistletoe adjuvant therapy for breast cancer, since the experimental group has a much higher risk of suffering from AEs. One important remark is that the great majority of the events were mild, within a range from 0.56 to 1.49% of discontinuation due to AE. Tolerance was rated as good by 94% of the patients [54] and by 78.9% of the physicians. [60] Overall, despite the higher chance of having an AE, mistletoe treatment would be regarded as safe.

Based on 6 and 2 observational studies respectively, quality of the evidence on survival and tumour response was rated down from low to very low because of the study limitations alone. The main sources of data for these two outcomes are the same studies by Grossarth-Maticek, showing the same methodological problems.

Due to lacking data (no data on control survival for 4 of the 6 studies), the measures of absolute effect could not be calculated for survival and tumour response hazard ratios. Tumour response assessment was also divided in four components because of the outcome measures used in the studies: local recurrences, lymphatic metastases, distant metastases and “all events” (including death). From these components, only data from local recurrence and distant metastases were included in the quality assessment, due to the non-fulfilment of the proportional hazard model adequacy at the lymphatic metastases and “all events” components.

Regarding QoL, even though evidence was looking promising with rating down for study limitations and indirectness being not necessary and for publication bias being questionable, it was impossible to derive a common measurement from the different scores used. EORTC sums up the scores from 0 to 4 for each of the questions (the higher the score, the worst the QoL), GLQ-8 is a visual analogue scale and FACT-G scale also sums the points for each question (but for two of the sections the higher the score the worst the QoL and for the other two sections the higher the score the better the QoL).

The same problem of different and incomparable outcome measurements made it also impossible to assess positive immunologic response and neutropenia.

Following GRADE guidelines, comments are included at the bottom of the table, with brief explanations on the reasons for the judgments.

Table 5. Final GRADE evidence profile table on adjuvant mistletoe treatment for breast cancer

Quality Assessment							Summary of findings						
							Nr. of Patients		Effect			Quality grading	
Nr. of Studies	Design	Limitations	Inconsistency	Indirectness	Imprecision	Publication bias	Mistletoe	Control	Relative	Absolute			
									Relative Risk 95% CI	Control Risk	Risk Difference 95% CI	Start grade	Final grade
Safety 2[1]	RCT	No serious limitations	No serious inconsistency	No serious indirectness	No serious imprecision	Possible[2]	378	246	8.94 (3.66 - 21.87)	20 per 1000	161 more per 1000 (54 - 424)	HIGH	HIGH[3]
Survival 6	Obs	Serious limitations[4]	No serious inconsistency	No serious indirectness	No serious imprecision[5]	Unlikely	1029	1051	0.46 (0.36 - 0.58)	Not estimable[6]	Not estimable[6]	LOW	VERY LOW[7]
Tumour response (Local recurrences) 2	Obs	Serious limitations[8]	No serious inconsistency	No serious indirectness	Serious imprecision[9]	Unlikely	122	122	0.42 (0.24 - 0.77)	Not estimable[6]	Not estimable[6]	LOW	VERY LOW[10]
Tumour response (Distant metastases) 2	Obs						122	122	0.41 (0.23 - 0.66)				
Quality of Life 3	RCT	No serious limitations	Not assessable[11]	No serious indirectness	Not assessable[12]	Possible	271	277	Not assessable[11]	Not assessable[11]	Not assessable[1]	HIGH	Not assessable[12]

1. For safety, only the RCT results were included, for they provided greater confidence in the estimates, following GRADE's guidelines (since observational studies were downgraded).
2. Even though publication bias was considered possible, the uncertainty about it led to the decision of not rating down for publication bias
3.The adverse effects were not specified, this constituted the only high risk at study limitations, thus not worth downgrading, but still limiting the evidence for the adverse effects.
Nevertheless, the higher rate of discontinuation due to AE was 1,49%, most AEs were mild and tolorability was regarded as good by 94% of patients (Beuth) and 78,9% of physicians (Bock).
4. One study with no serious limitations, but 5 lacking accuracy and precision of the data, missing data on treatment regimes, some prognostic factors not included
5. Estimation of OIS not exact, but even worst estimation still met OIS
6. Calculations not possible with data disclosed within articles, limitation of not asking authors for needed missing data
7. Rated down for limitations
8. Studies lacking accuracy and precision of the data, missing data on treatment regimes, some prognostic factors not included
9. Optimal Information Size not met
10. Rated down for limitations and imprecision
11. Not possible to perform calculations, since no common score was used and the differences between them do not allow a transformation of results.
12. Since assessments on inconsistency and imprecision were not possible, it is also not possible to know if the quality would have been downgraded or not.

Discussion

This review had the purpose of investigating what would the evidence on mistletoe for breast cancer look like when disclosed according to the GRADE process and if GRADE would be appropriate for evaluating CAM evidence.

The main findings led to the conclusions that GRADE method is adequate for assessing CAM evidence, because it enables the inclusion of observational studies in the process, and performs a transparent assessment of all study designs. Nevertheless, the system is labour-intensive and still relies on subjective interpretation of some of the criteria.

Regarding the evidence on mistletoe, this assessment showed high quality evidence on safety and presumably on QoL and low quality on evidence for survival and tumour response, but all pointing to positive effects.

Study search: availability, German language, limitations

Within the publication selection process, 53 articles were excluded because of having being written in another language rather than English and because most of them were also not available, not even as abstract.

Investigating the 10 available articles or abstracts in German among those articles, even if they would have been included in the analysis, the contribution would not be significant, since 5 are doubles from articles already included in the analysis, one is a case report, one is a qualitative study on 4 patients (incomparable outcome measure), one is a feasibility study on immunologic parameters and EORTC that found no significant differences between test and control groups, and one only described difficulties enrolling patients for mistletoe studies. It is not clear if one of the studies could have added to the safety assessment, since no reference is made on the control group in the abstract.

Another limitation is that the search was performed only with English terms, and by doing so negative studies published in local journals or as “grey literature” might have been missed. Furthermore, authors were not contacted for studies that were not available online (30 studies), which means that not even the abstracts were available. Thus, the sample of included studies might not be representative, and the estimates of effect could possibly be overestimated by the absence of negative studies, or underestimated by lacking significant studies written in other languages.

Remarks when specifically assessing quality – experience of using GRADE

In order to refine its process and address areas of uncertainty, the GRADE approach is in constant development, therefore some of the difficulties described below might be addressed in future adaptations. The authors also recognize that not only methodological advances are expected to be developed and thus included in the GRADE approach, but also the revision of established concepts.[38]

Many of the institutions adopting GRADE value the advantage of GRADE separating the quality of the supporting evidence from the strength of the recommendation. This permits strong recommendations endorsed by low-quality evidence from observational studies as well as weak recommendations sustained by high-quality evidence. Since this analysis of the evidence on mistletoe and breast cancer was performed only as a systematic review, the recommendation step in the GRADE process was not included. Most of these institutions are also of the opinion that its use helped securing transparency, consistency and systematic approach when assessing evidence. Furthermore, elucidating the weight of the available evidence and

highlighting low quality evidence more easily, as well as enabling problems and questions to be raised and properly discussed were also appreciated qualities. [44,61,62,63,64]

One particular challenge described by the GRADE organizers and also experienced in this review is the uncertainty of the weight to put when judging the different criteria and when downgrading. GRADE acknowledges that different groups analyzing the same evidence might come up with different results, depending on their view of what is relevant or on their "generosity" [50,51] but even equally competent and honest reviewers might have disagreement on interpretation of evidence.[44] So, although the assessment is explicit and transparent, subjectivity plays an important role. [45] Furthermore, evaluators' research experience and knowledge, in addition to familiarity with the intervention and outcomes, also play fundamental roles, influencing one's capability of knowing what is relevant and what is not in a specific context or for a specific outcome. In this sense, this review has also its limitations, since the main researcher is a physician, familiar with cancer and mistletoe therapy, but a junior researcher.

When more than two interventions are being compared, the usual pairwise analysis can result in one evidence profile table for each pair. Until now, GRADE offers no clear framework for the synthesis of the overall interpretation of all the profiles, as well as for presenting network meta-analysis, but this is already being developed by the GRADE group. [63]

Quality criteria

There were specific difficulties when going through each criterion.

When appraising **study limitations** (risk of bias), although the elements under scrutiny are fewer and somehow more flexible than the ones observed in a Cochrane review, the limited evidence supporting these criteria [45] made this judgment somehow questionable.

When assessing **inconsistency**, one of the four criteria - I^2 - is derived from the same calculations for statistical test for heterogeneity (p-value). Since they "point in the same direction", maybe they should be valued together as a single criterion.

Calculating OIS (optimal information size) when the relative measures available are HRs is more difficult than for RR, and all data necessary for performing these calculations may not always be available in the articles, such as control median survival, length of the time to recruit patients for study to be discounted from the follow-up duration. As already mentioned in the results, missing data not obtained with the authors restricted the evaluation. Also the reasoning on how to consider the CI boundaries depends on judgments on values and preferences, which will be influenced by the importance of the outcome, the adverse effects, intervention's practicability, even resource use. Thus, setting a clinical decision threshold requests also a lot of professional experience in the topic, and was also not performed in this analysis, since it was not meant for guideline development.

Confidence in the presence or absence of **publication bias** is mentioned by GRADE organizers as somewhat difficult, as well as placing a threshold for rating down for it, what was also experienced at this review (GRADE 5). [49] In this case, since mistletoe is a therapy mainly prescribed in German speaking countries, most research on the topic is also carried out in these countries, or by German/Swiss researchers, and so negative results might have been published in local journals that will not appear in a search done with English words, as already explained previously. In this review, double counting was carefully removed.

Industry sponsorship was also another characteristic that was challenging to judge. Many studies had the medication provided by pharmaceutical industries, but were performed by academic hospitals or independent research centra that claimed no conflict of interest. It was unclear whether this would be reason enough to

rate the evidence down for publication bias. Future studies could at least make a clear statement that the sponsor did not interfere in study planning, design, conduct or analysis. [24]

Regarding funnel plots, since there is evidence and debate on the non-reliability of its asymmetry (statistically tested or visually interpreted) in predicting publication bias, especially for a small number of articles, and since this review dealt with less than 10 studies per outcome, they were not considered when assessing publication bias.[65]

Even though GRADE allows recommendations being made even with low quality evidence, and before any change happens within the current prevailing EBM scientific paradigm, CAM researchers will be able to have the body of evidence provided by observational studies upgraded to moderate and even high quality if the criteria of large magnitude of effect, dose-response gradient or effect being possibly reduced by confounders are met.[42] But since only methodologically rigorous studies can be upgraded, their methodological quality has to be improved.

Outcomes

The assessment of each outcome also brought some challenges.

In safety, although articles with no control groups had already been excluded in the search, the quality assessment of the observational studies showed many problems. While having a control group for the investigation of other outcomes in the studies, such as survival or QoL, these studies and one of the RCTs (Tröger) showed no surveillance of adverse effects in the control group and consequently, no data on that.

Furthermore, the description of AEs varied between studies and even within studies. There were no clear definitions of which were considered systemic or local AEs, definitions described at methods section differed from the definitions mentioned at results section, there was little information on the mode of data collection, timing, attribution methods, intensity of ascertainment, and harm-related monitoring.

No reference was made to confounding factors such as the possibility of some of the AEs related to mistletoe being symptoms from the disease itself, from the standard conventional treatment that both groups were receiving equally or even from the injection per se. Fever is mentioned as one of the possible mistletoe related AEs, but there is plenty of evidence that it might be one of the symptoms of breast cancer or treatment, as well as headache, cold and flu symptoms, also described as possible mistletoe AEs. Injection-site reactions can be caused by the mistletoe medication, but also from glass particles from the ampoule, by infection, histaminic reaction to the needle, poor technique in application.[66,67,68,69,70,71]

Blinding is also frequently said to be impossible when investigating mistletoe, since the typical reactions cannot be imitated by a pseudo-placebo and thus would unblind the patient and the assessing researcher [53,72,73]. The studies included in this review showed a range of AEs presumably related to mistletoe from 10% to 32.4% for the experimental groups. In the study with the highest adverse reaction rate, in 67.6% of the treated population blinding was still functional. Not to mention that the studies that had controls using placebo injections had injection-site reactions ranging from 1.7 to 2.8%, a fact that may also speak in favour of blinding mistletoe studies being at least worth of further discussion.

The safety assessment of this review showed that, despite being frequent, AEs related to mistletoe therapy are mostly mild, well tolerated and with spontaneous remission, as also shown in the 2006 review of anthroposophic medicine's effectiveness, utility, costs and safety. In this HTA assessment, mistletoe treatment studies included more than 10,000 patients, with local and systemic reactions and tolerability reportedly similar to the ones found in this review.[16]

Since safety evaluation is critical to any intervention assessment, a more uniform AE assessment could improve the quality of the evidence supporting safety. In this sense, although meant for RCTs, the CONSORT extension for reporting of harm-related data presents 10 recommendations that could be considered in future studies, as well as the Common Terminology Criteria for Adverse Events (CTCAE), and MedDRA - the Medical Dictionary for Regulatory Activities. Another point worth of discussion for future studies is weather AEs should be counted as 1 for each patient that showed AEs, independently on the number of events per patient, or on the medication dose (especially in the dose-increasing trials). Perhaps the rate of AEs per number of injections per dose would be more accurate.[74,75,76,77]

In the quality appraisal of survival evidence, studies presented each experimental group compared to a control group at the same cancer stage. The HR range was from 0.27 to 0.65 regardless of the cancer stage, and more advanced stages (such as lymphatic metastases) had smaller HR (0.27, CI: 0.15 to 0.5) when compared to local cancer (0.65, CI: 0.34 to 1.25). Thus the test whether the effect on survival would be positive regardless of the stage, despite having been performed on only 6 studies, revealed that any stage benefits from the treatment regarding survival. The pooled HR was 0.46 (CI: 0.36 to 0.88).

The 22 studies comparing mistletoe to no extra treatment in a 2009 systematic review on survival (all cancers, not only breast) had an overall HR estimated in 0,59 (CI: 0.53 to 0.66) and found that tumour localization was not significantly associated with better or worse study outcome. The methodological quality of the studies was also heterogeneous, and even though quality has been improving in the most recent studies, problems of missing data on compliance, follow-up, ITT analysis, dose and frequency of mistletoe treatment were found, as in this review. Despite the unique obstacles that research in this area might face, poor quality was predominantly driven by low levels of documentation quality. Randomized studies showed less effect than non-randomized studies and significantly better results were shown by matched-pair studies. In this review most studies were matched – paired, so if RCTs would have been available for survival in breast cancer patients treated with mistletoe, eventually a smaller effect could have been shown in an extra line at the evidence profile table.[24]

In a review on the effect of mistletoe on breast and gynaecological cancers by Kienle et al, from the 4 RCTs included, two showed statistically significant benefit on survival and two a positive trend. From the three non-randomized studies, two reported statistically significant benefit and one a small positive trend.[23]

Tumour progression/response showed similar HRs in the two studies (0.42, CI: 0.24 to 0.77 and 0.41, CI: 0.23 to 0.66), but lacking data on control group (same as survival analysis) prevented the absolute measures calculations from being done. Tröger's randomized study, mentioned at the safety and QoL assessments, had a continuation, and its results were published after the search for this review was performed. During a follow-up of 5 years, without any further mistletoe or conventional treatments, an annual visit documented the occurrence of relapse and/or metastases in both the experimental and control groups. No significant difference was found in the disease-free 5-year survival between mistletoe and control groups, not even between radiotherapy and hormonal therapy subgroups. This speaks in favour of mistletoe not having any detrimental effect on the conventional treatment, but also not having any advantage regarding disease-free survival.[78] If a comparison could be made with the survival data mentioned above, where RCTs showed less effect than non-randomized studies and that matched-pair showed bigger effects, this could be a reason why the study by Tröger showed smaller effects than the ones by Grossarth-Maticek.

In Kiene's 2009 review, of the 4 controlled studies that combined mistletoe and conventional breast cancer treatment, none found a disadvantage regarding disease recurrence and time to disease relapse.[23]

Appraisal of evidence on QoL proved to be more complex than expected, due to the different instruments chosen by each study to measure it, and their respective difference in units, but also in the direction of the results (better or worse QoL), as previously described in the results. Analysing the scales simultaneously, a

statistically significant difference in favour of the experimental group that was common to the three studies could be found for reduction of anxiety/depression or emotional well-being, less nausea/vomiting and less fatigue.

This is in line with the 2010 systematic review of controlled clinical studies on the influence of mistletoe on QoI in cancer patients (not only breast cancer) by Kienle and Kiene. From the 4 double-blind RCTs within the 36 included studies, 3 indicated a significant benefit, and a small pilot trial found no difference. Most benefits could be observed in mistletoe treatment concurrent with conventional cancer treatment, with better tolerability of the latter. Type or stage of disease did again not interfere and most positive results were reached with breast cancer patients. Feasibility of blinding in mistletoe studies was also mentioned, as well as whether an insufficient blinding could provide more valid results than an open application. Also which interference this could possibly have when assessing the placebo effect of obliging reporting (participants answering what they believe is wanted) is difficult to evaluate.[18]

In the 2009 breast and gynaecological cancer review, 19 of the 21 studies (11 RCTs, 6 non-randomized and 4 single-arm studies) reported benefit regarding QoL, mostly significant.[23] A further 2012 meta-analysis by Büssing on the same topic also concluded that the 16 included studies were of poor quality and small size, and showed that randomization studies did not differ from non-randomized ones in the multivariable metaregression. All publications investigating the effects of mistletoe applied to QoL found positive effects, but subject to bias that could contribute to the overall positive effect.[25]

Difference in evidence quality between outcomes

As mentioned by the GRADE group, there can be difference in quality of evidence between outcomes.[39]

In this review, this could be confirmed for some cases. The RCT by Tröger had limitations that were differently judged in safety and QoL assessment. Lack of blinding was considered as high risk in safety, due to the consequent evaluation of AEs, but was regarded as low risk in QoL, because questionnaires were answered by the patients themselves and not by the researcher (obliging reporting could still be a consequence of non-blinding). Selective outcome reporting was also considered high risk for the same study in safety, due to differences in the description of the AEs in the methodology and result section and no assessment of AEs in the control group. In Qol, the same criterion was judged as low risk, because the outcome measure was adequately described and reported for both groups.

The same differences could be found in the 2006 Semiglasov study for both outcomes and the same criteria, but the limitations in safety were not so important that they demanded rating down the evidence or excluding the study from the analysis, as it was the case for the study by Tröger.

The study by Bock also performed better in QoL when compared to safety, since the control group was not included in the latter, just as the outcome measurement was flawed (no clear description of the AEs, measurement of systemic and local reactions together, confounding was not controlled for), and in the former these limitations were absent.

GRADE and complex interventions

Some reflections should perhaps be presented on the suitability of the current evidence-based hierarchical view (and the quality assessment systems based on it) to the investigation of complex interventions. Surgery, rehabilitation, nursing, psychotherapy, more complex interventions of CAM, public health interventions, among others, find difficulties fitting the RCT-hegemonic model when designing their studies.

A review of the application of GRADE by 25 groups in the field of public health (the majority within the WHO) described challenges that might also be common to other complex interventions.[79] For example, most studies assessing complex interventions focus on the impact of the whole package, rather than on each component of the intervention, so PICO questions have to consider the intervention as a whole or stress a presumed active component. Both cases will interfere with the selection of studies.

The choice of outcome and outcome measures is another experienced difficulty, since complex interventions have multiple outcomes, also at the individual and group levels, thus having implications for indirectness when grouping them or dealing with the variations in assessment scales.

GRADE also does not allow non-epidemiological evidence such as mechanistic, animal or laboratory evidence, rationales from other disciplines such as engineering, physiology, chemistry, physics, to be integrated in the assessment of a body of evidence. In fact, GRADE argues that evidence coming from animal studies should be downgraded two levels for indirectness.[48,79]

For interventions already regarded as safe, when the mechanism of action is not measurable (case of many CAM therapies), or also in case of complex interventions, economic evaluations/cost-effectiveness analyses are a possible start to assess advantages. GRADE article on resource use (guidelines number 17) is still to be published, but the necessity to tackle this topic was already mentioned by The Endocrine Society when developing its clinical practice guidelines, due to challenges on the clarity, validity, applicability, and interpretability of the evidence coming from economic analyses.[62]

GRADE restricts the quality analysis to benefits and harms of interventions and still does not include other relevant factors also to be valued like disease burden, availability of the intervention, cost-effectiveness, feasibility issues, modelled estimates. This limitation was mentioned by the group developing WHO's immunization recommendations using GRADE, and also by the group appraising GRADE applied to public health interventions.[64,79] NICE came up with its own solution and presented resource use as one of the outcomes in an adapted economic evidence profile table showing the study, limitations, applicability, other comments, incremental cost, incremental effects, ICER and uncertainty in the ICER estimate.

Conclusion

This review, through the GRADE approach, assessed the quality of the evidence provided by studies on adjuvant mistletoe treatment for breast cancer. The quality proved to be high for safety studies (RCTs) and very low for survival and tumour response (observational studies), which is in line with evidence from other reviews. The initial quality of QoL studies was regarded as high, but due to studies' particularities, the final assessment was not possible using the GRADE system.

In general, most of the downgrading of the evidence was due to study limitations/risk of bias. Improvement in this aspect could help ameliorating the overall quality of the evidence on mistletoe and breast cancer and would enable the eventual upgrading of the quality of methodologically rigorous observational studies, providing better evidence for its recommendation.

GRADE proved to be suitable for assessing evidence coming from CAM studies for the innovations it proposes. Besides being an organized and transparent way of assessing quality of evidence, it deals with many aspects differently from previous methods, like considering evidence from observational studies and assessing quality per outcome. This makes it a good option for assessing evidence on CAM, often consisting of more observational studies than RCTs.

Even though the GRADE method has some critics from institutions that have already been using it, it offers an organized and transparent way of assessing quality of evidence. Since it is not a static approach, further developments, improvements and details of usage are expected.

Appendices

Appendix 1. Standardized quality assessment tables – SAFETY

Table 1. Quality assessment for study limitations (risk of bias) for observational studies

Observational												
Study	**Design**	**Limitations (risk of bias / internal validity)**										
		Failure to develop and apply appropriate eligibility criteria (inclusion of control population)	Under- or overmatching in case-control studies	Selection of exposed and unexposed in cohort studies from different populations	Flawed measurement of both exposure and outcome	Differences in measurement of exposure	Differential surveillance for outcome in exposed and unexposed in cohort studies	Failure to adequately control confounding	Failure of accurate measurement of all known important prognostic factors	Failure to match for prognostic factors and/or lack of adjustment in statistical analysis	Incomplete follow-up	Result
Bock 2004	Retrolective Comparative Cohort (parallel group design)	yes = no inclusion of control group in the adverse events assessment.	no	no	yes = no clear description of adverse events (only mentions common toxicity criteria and separates by local and systemic drug adverse reaction only), no separate presentation of the different kinds of adverse events	yes = adverse events not measured in control group	yes = no surveillance for adverse effetcs in control group	yes = confounding for mistletoe common adverse events like fever/chills not considered	no - tables 1 and 2 (pg 27) - quite complete, test group even showed more severe and advanced disease	no - multivariate adjusting and stratification	**intention-to-treat analysis** + 10.3% excluded for severe protocol violation (sensitivity analysis did not detect significant bias on the adjusted outcomes due to exclusion)	Very serious limitations
Beuth 2008	Retrospective Comparative Cohort	yes = no inclusion of control group in the adverse events assessment.	no	no	yes = no clear description of adverse events, no separate presentation of the different kinds of adverse events	yes = adverse events not measured in control group	yes = no surveillance for adverse effetcs in control group	yes = confounding for mistletoe common adverse events like fever/chills not considered	no - page 525, test group showed a tendency to more advanced disease (higher UICC stages) and more frequent hormone-therapy in the control group (advantages for control group)	no - logistic regression analysis	2.4% study group and 1.8% control group (died)	

Table 2. Quality assessment for study limitations (risk of bias) for randomized controlled trials

RCT													
Study	**Design**	**Limitations (risk of bias / internal validity)**											
		Lack of allocation concealment	Lack of blinding	Incomplete accounting of patients and outcome events	Selective outcome reporting bias	Stopping early for benefit	Use of unvalidated outcome measures	Carryover effects in crossover trial	Recruitment bias in cluster-randomized trials	Baseline values	Unappropriate consideration of Intention-to-treat principle	Conclusion	Result
Tröger 2009	Prospective randomized open label clinical trial (pilot study)	Low risk. Probably done. Quote: "Allocation concealment was implemented by using sealed envelopes" (...)	High risk. Probably not done. Quote: " CRDT (Clinical Research Dr. Tröger) was responsible for planning, conduct, monitoring, and analysis of the study." Reason for not placebo-controlled: "typical reactions following injections of mistletoe extracts cannot be imitated by a pseudo-placebo".	Low risk. One patient control group withdrawn for heart disease (3%); less than 1% missing data at both groups for both assessed parameters	High risk. Planned outcomes reported, but methods section suggests that other adverse events then injection site reaction were measured, but no information on this issue (i.e. flu-like symptoms), no results presented for adverse events in control group.	Low risk.	Low risk (absent).	Low risk (absent).	Absent	Balanced (pg 38)	Absent (pg37: "statistical analysis was performed on the intention-to-treat population").	Crucial limitation for two criteria.	Excluded from analysis
Semiglazov 2006	RCT (double blind)	Low risk. Probably done. Quote: "Patients were allocated to the treatment groups on the basis of a computer-generated randomisation list."	Low risk. Double blind.	Low risk. 4 patients lost to placebo group = 2% (2 decision of the patient, 1 serious adverse effect, 2 patients moved). 1 patient lost to test group = 0,6% (decision of patient)	High risk. Planned outcomes reported, but adverse events and safety laboratory tests not specified at methods section.	Low risk.	Low risk (absent).	Low risk (absent).	Absent	Balanced between treatment groups with respect to demographics and medical history, not with respect to QoL scale scores (stronger restrtiction at test group) and histological classification (lower rate of invasive tumors inplacebo group), evaluation complemented by adjusted analyses of variance.	All patients included in the evaluation.	High risk of bias for one criterion	No serious limitations
Semiglasov 2004	RCT (double blind)	Low risk. Probably done. Quote: "Patients were allocated to the treatment groups on the basis of a computer-generated randomisation list."	Low risk. Double blind.	Low risk. 4% (11 patients lost: 4 to adverse effects, 4 decision of patient, 3 other reasons) ITT analysis.	High risk. Planned outcomes reported, but adverse events and safety laboratory tests not specified at methods section.	Low risk.	Low risk (absent).	Low risk (absent).	Absent	Balanced between treatment groups with respect to demographics and medical history, not with respect to QoL scale scores (not statistically significant)	Absent (ITT analysis performed).	High risk of bias for one criterion	

Table 3. Quality assessment for inconsistency and imprecision (RCTs)

Study	Design	Inconsistency - In relative, not absolute, measures of effect (RR/HR/OR) - Rate down if:					Imprecision (around absolute, rather than relative effects)		
		Point estimates vary widely across studies	Confidence intervals (Cis) show minimal or no overlap	Statistical test for heterogeneity shows a low p-value (<0.10 suggested by Cochrane for few studies instead of <0.05)	I^2 is large	**Result**	If the OIS (optimal information size) is not met, rate down for imprecision unless sample size very large.	95% Confidence Interval excludes no effect (CI around RR excludes 1.0)	**Result** If OIS met and 95% CI around effect excludes 1.0, no need to rate down for imprecision.
Semiglazov 2006	RCT (double blind)	no	Confidence intervals show reasonable overlap	p = 0.7052	I^2 is zero	**No serious inconsistency**	OIS met (calculated OIS = 56)	No effect excluded	**No serious imprecision**
Semiglasov 2004	RCT (double blind)								

	ln(RR)	SE(ln(RR))	wi = 1/VAR	Q	Δ2	wi*		I-square
Semiglasov 2004	1.9847	0.710333242	1.981873	0.1431505	0.000000000	1.9819		-599%
Semiglasov 2006	2.3354	0.595170363	2.823042			2.8230		
								Q: Chi test
								0.70517
				Pe_s	**random effects**			
				2.1908	PeDL	SE of PeDL		
					2.19075	0.45620	95% CI	
				RR	8.94	1.578069019	3.66	21.87

Study	Relative risk	ln RR	SE of the log RR	95% CI ln(RR)		95% CI RR	
Semiglasov 2006	10.3333	2.335374916	0.595170363	1.168841004	3.501908828	3.218260524	33.17872403
Semiglasov 2004	7.2772	1.984749982	0.710333242	0.592496828	3.377003136	1.808498293	29.28288268

Calculating Relative Risk

Semiglasov 2006		Adverse effects	No adverse effects	Total		**SE of the log RR**
	Test	31	145	176		0.595170363
	Control	3	173	176		
	Total	34	318	352		
	Relative Risk	Test Event Rate	0.1761	10.3333		
		Control Event Rate	0.0170			

Semiglasov 2004		Adverse effects	No adverse effects	Total		**SE of the log RR**
	Test	42	160	202		0.710333242
	Control	2	68	70		
	Total	44	228	272		
	Relative Risk	Test Event Rate	0.2079	7.2772		
		Control Event Rate	0.0286			

Figure 1. Forest plot

Study or Subgroup	Experimental Events	Experimental Total	Control Events	Control Total	Weight	Risk Ratio M-H, Random, 95% CI
Semiglasov 2004	42	202	2	70	41.2%	7.28 [1.81, 29.28]
Semiglasov 2006	31	176	3	176	58.8%	10.33 [3.22, 33.18]
Total (95% CI)		**378**		**246**	**100.0%**	**8.94 [3.66, 21.87]**
Total events	73		5			

Heterogeneity: Tau² = 0.00; Chi² = 0.14, df = 1 (P = 0.70); I² = 0%
Test for overall effect: Z = 4.80 (P < 0.00001)

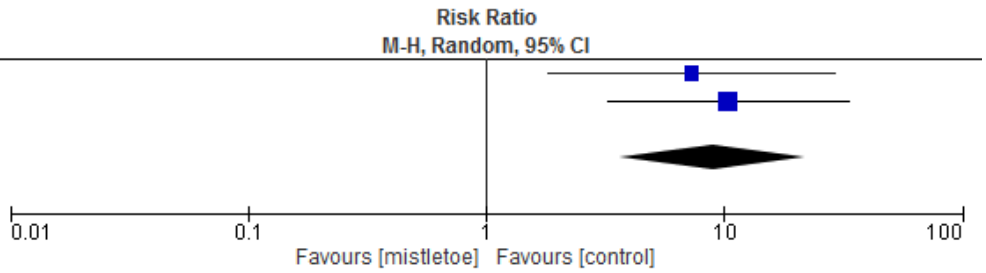

Table 4. Quality assessment for inconsistency and imprecision (observational studies)

Observational

Study	Design	Inconsistency - In relative, not absolute, measures of effect (RR/HR/OR) - Rate down if:					Imprecision (around absolute, rather than relative effects)		
		Point estimates vary widely across studies	Confidence intervals (Cis) show minimal or no overlap	Statistical test for heterogeneity shows a low p-value (<0.10 suggested by Cochrane for few studies instead of <0.05)	I^2 is large	**Result**	If the OIS (optimal information size) is not met, rate down for imprecision unless sample size very large.	95% Confidence Interval excludes no effect (CI around RR excludes 1.0)	**Result** If OIS met and 95% CI around effect excludes 1.0, no need to rate down for imprecision.
Bock 2004	Retrolective Comparative Cohort (parallel group design)	Calculations not possible (no control group data). 18.1% of the test patients at Bock and 10% at Beuth experienced adverse drug reactions presumably related to mistletoe, with respective good tolerance of 78,9% and 94%.				Not assessable	Calculations not possible (no control group data)		Not assessable
Beuth 2008	Retrospective Comparative Cohort								

Table 5. Quality assessment for indirectness and publication bias

Study	Design	Indirectness						Publication Bias					
		Differences in population (applicability)	Differences in interventions (applicability)	Differences in outcome measures (surrogate outcomes)	Indirect comparisons because of biased head-to-head comparisons (e.g. industry)	Mechanism of action	Result	Early positive studies, small in size, preliminary and pilot studies	Non-English speaking countries might submit negative studies to local journals, "gray literature"	Double counting	Industry sponsored (or likely to be) or conflict of interest	Assymetrical funnel plot	Result
Observational													
Bock 2004	Retrolective Comparative Cohort (parallel group design)	not the case	not the case	not the case	not the case	not the case	no serious indirectness	No	Possible	not the case (revised for dupplicates)	All studies performed by research institutes (public, universities, Bock=independent research institute).	not possible	Possible Publication bias
Beuth 2008	Retrospective Comparative Cohort												
RCT													
Semiglazov 2006	RCT (double blind)	not the case	not the case	not the case	not the case	not the case	no serious indirectness	Not early positive, Semiglasov 2004 pilot.	Possible	not the case (revised for dupplicates)	Both received funding/medication provision from Madaus GmbH, but were performed by independent research centra.	Not considered (too few studies)	Possible Publication bias
Semiglasov 2004	RCT (double blind)												

Table 6. Evidence profile table for safety (adverse events)

Quality Assessment							Summary of findings						
							Nr. of Patients		Effect			Quality grading	
Nr. of Studies	Design	Limitations	Inconsistency	Indirectness	Imprecision	Publication bias	Mistletoe	Control	Relative	Absolute			
									Relative Risk 95% CI	Control Risk	Risk Difference 95% CI	Start grade	Final grade
2	Obs	Very serious[1]	Not assessable[2]	No serious indirectness	Not assessable[2]	Possible[3]	877	1246	Not assessable[2]	Not assessable[2]	Not assessable[2]	LOW	VERY LOW[4]
2	RCT	No serious limitations	No serious inconsistency	No serious indirectness	No serious imprecision	Possible[3]	378	246	8.94 (3.66 - 21.87)	20 per 1000	161 more per 1000 (54 - 424)	HIGH	HIGH[5]

1. No clear description of adverse events, no measurement of AEs in control group, confounding with cancer symptoms not considered.
2. Control group not included in the adverse event measurement.
3. Even though publication bias was considered possible, the uncertainty about it led to the decision of not rating down for publication bias
4. Rated down for limitations. Since it has already reached the lowest quality, even if downgrading for inconsistency and imprecision would be possible, it would not make any difference.
5.The adverse effects were not specified, this constituted the only high risk at study limitations, thus not worth downgrading, but still limiting the evidence for the adverse effects.

Control Event Rate	0.020
Test Event Rate	0.182
Risk Difference	0.161
95% CI Risk Difference	0.054 to 0.424

Appendix 2. Standardized quality assessment tables – SURVIVAL

Table 1. Quality assessment for study limitations (risk of bias)

Observational												
Study	**Design**	**Limitations (risk of bias / internal validity)**										
		Failure to develop and apply appropriate eligibility criteria (inclusion of control population)	Under- or overmatching in case-control studies	Selection of exposed and unexposed in cohort studies from different populations	Flawed measurement of both exposure and outcome	Differences in measurement of exposure	Differential surveillance for outcome in exposed and unexposed in cohort studies	Failure to adequately control confounding	Failure of accurate measurement of all known important prognostic factors	Failure to match for prognostic factors and/or lack of adjustment in statistical analysis	Incomplete follow-up	Result
Bock 2004	Observational (comparative cohort with parallel groups)	no (pg26)	no	no	no - standardized case report forms (CRFs) tested for comprehension, plausibility and reliability, filled in by trained personnel +quality control	no - standardized case report forms (CRFs) tested for comprehension, plausibility and reliability, filled in by trained personnel +quality control	no (retrolective and data collected from the same oncologic centres)	no - page 28, also adjusting for treatment regimen	no - tables 1 and 2 (pg 27) - quite complete, test group even showed more severe and advanced disease	multivariate adjusting and stratification	no	Serious limitations
Grossarth-Maticek 2004 A	Observational (cohort) - prospective randomized matched-pair study **(MammaRand)**	Inclusion and exclusion criteria well developed and applied (pg 287), but due to study commencement in 1968, no written protocol, no pre-specified formulation of statistical hypotheses, no sample-size calculations in advance.	no	no	Even though predetermined case report forms were used, checked for consistency, missing data, and independent reviews on overall quality were performed (pg 288), authors recognize that accuracy and precision of the data are poor (exact dates of first diagnosis, operation, some data assessments and matching not available). Since this applies to both groups it would not affect one more than the other, but affects the overall quality of the results (not accurate).	no - predetermined case report forms, checked for consistency, missing data, ondependent reviews on overall quality (pg 288)	no	No data on doses, variation in dose, breaks in therapy, host trees etc (mistletoe therapy was prescribed by attending doctors, not study physicians).	Some medical prognostic factors were either not recorded throughout all cases, or not recorded at all (i.e. steroid receptor; histopathological type), performed matching for hormone therapy could be a proxy for hormone receptor.	no -adequate match for the considered prognostic factors (tables 3 and 4 of supplemental material)	21 patients (6 from the treatment group) dropped out of the study (27%), so 21 matched-pairs were eliminated (from the 38). Does not affect randomization (performed pairwise and both partners were eliminated), but very small size. Patients were included in the final analysis (ITT).	
	Observational (cohort) - non-randomized matched-pair study **(Mamma)**										8% attrition bias. Impairments of internal validity due to drop-outs neutralized by exclusion of the corresponding match. All patients included in the complete set analyses.	
Grossarth-Maticek 2004 B	Observational (cohort) - Randomized matched-pair study **(MammaLymRand)**	Inclusion and exclusion criteria well developed and applied (pg 287), but due to study commencement in 1968, no written protocol, no pre-specified formulation of statistical hypotheses, no sample-size calculations in advance.	no	no	no - case report forms	no	no - equal analysis	No data on doses, variation in dose, breaks in therapy, host trees etc (mistletoe therapy was prescribed by attending doctors, not study physicians) Assumed subcutaneously, 2-3 times/week for being the usual approach.	Some medical prognostic factors were either not recorded throughout all cases, or not recorded at all (i.e. steroid receptor; histopathological type), performed matching for hormone therapy could be a proxy for hormone receptor.	no - matching "almost perfect"	no loss to follow-up	
	Observational (cohort) - non-randomized matched-pair study **(MammaRec)**									no - sensitivity analysis showed original sets were fairly balanced regarding prognostic factors	One patient of each group lost follow-up, not	
	Observational (cohort) - non-randomized matched-pair study **(MammaLym)**										One control patient lost follow-up	
	Observational (cohort) - non-randomized matched-pair study **(MammaMet)**										3 control patients lost follow-up	

Table 2. Quality assessment for inconsistency and imprecision

Study	Design	**Inconsistency** - In relative, not absolute, measures of effect (RR/HR/OR) - Rate down if:					**Imprecision (around absolute, rather than relative effects)**		
		Point estimates vary widely across studies	Confidence intervals (CIs) show minimal or no overlap	Statistical test for heterogeneity shows a low p-value (<0.10 suggested by Cochrane for few	I^2 is large	**Result**	If the OIS (optimal information size) is not met, rate down for imprecision unless sample size very large.	95% Confidence Interval excludes no effect (CI around HR excludes 1.0)	**Result** If OIS met and 95% CI around effect excludes
Bock 2004	Observational (comparative cohort with parallel groups)	Point estimates do not vary widely across studies	Confidence intervals (CIs) show reasonable overlap	0.4743	I^2 is zero (negative values are put equal to zero)	No serious inconsistency	Exact OIS impossible to calculate due to lack of data from all studies, but all estimations calculated with available data show OIS met	Pooled HR excludes 1.0	No serious imprecision
Grossarth 2006 A Mamma	Observational (cohort) - non-randomized matched-pair								
Grossarth 2006 A MammaRand	Observational (cohort) - prospective randomized matched-pair								
Grossarth 2006 B MammaRec	Observational (cohort) - non-randomized matched-pair								
Grossarth 2006 B MammaLym	Observational (cohort) - non-randomized matched-pair								
Grossarth 2006 B MammaMet	Observational (cohort) - non-randomized matched-pair								

	ln(HR)	SE(ln(HR))	wi = 1/VAR	Q	Δ2	wi*		I-square
Bock 2004	-0.78	0.38	6.93	4.542	0.00000000	6.93		-10%
Grossarth 2006 A Mamma	-0.84	0.24	17.36			17.36		Q: Chi test 0.47430
Grossarth 2006 A MammaRand	-0.43	0.33	9.18			9.18		
Grossarth 2006 B MammaRec	-0.65	0.41	5.95			5.95		
Grossarth 2006 B MammaLym	-1.31	0.31	10.41			10.41		
Grossarth 2006 B MammaMet	-0.63	0.26	14.79			14.79		
				Pes	random effects			
				-0.79	PeDL	SE of PeDL		
					-0.79	0.12	95% CI	
				HR	0.46		0.36	0.58

OIS Calculation (only data from Grossarth-Maticek available, for rough estimation, also worse scenarios tested within data across studies, all fitted the OIS)				
http://www.cct.cuhk.edu.hk/stat/survival/Rubinstein1981.htm				
α (significance level)	0.05			
β (1-β = power of the test)	0.2	Best estimation	N = 214 (107 control and 107 experimental)	
δ (hazards ratio)	0.65	Worst estimation	N=787 (398 control and 389 experimental)	
Ms (control median survival)	143.7			
Qc (proportion control group)	0.5			
Qe (proportion experimental group)	0.5			
T0 (recruitment time in months)	36			
T-T0 (follow-up duration in months)	252			

Figure 1. Forest plot

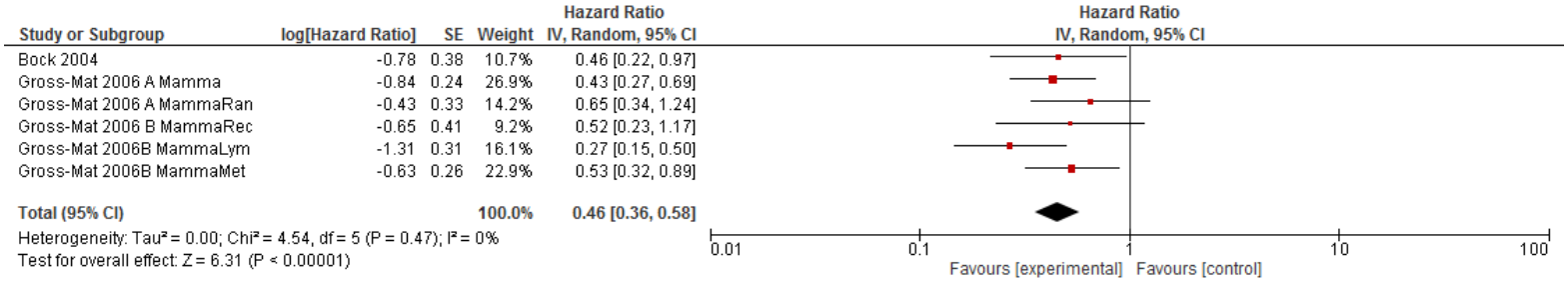

Table 3. Quality assessment for indirectness and publication bias

Study	Design	Indirectness						Publication Bias					
		Differences in population (applicability)	Differences in interventions (applicability)	Differences in outcome measures (surrogate outcomes)	Indirect comparisons because of biased head-to-head comparisons (e.g. industry)	Mechanism of action	Result	Early positive studies, small in size, preliminary and pilot studies	Non-English speaking countries might submit negative studies to local journals, "gray literature"	Double counting	Industry sponsored (or likely to be) or conflict of interest	Assymetrical funnel plot	Result
Bock 2004	Observational (comparative cohort with parallel groups)	not the case	not the case	not the case	not the case	not the case	No serious indirectness	Unlikely	Possible	no	not the case	Not considered (too few studies)	Unlikely
Grossarth-Maticek 2004 A	Observational (cohort) - prospective randomized matched-pair study **(MammaRand)**												
	Observational (cohort) - non-randomized matched-pair study **(Mamma)**												
Grossarth-Maticek 2004 B	Observational (cohort) - Randomized matched-pair study **(MammaLymRand)**												
	Observational (cohort) - non-randomized matched-pair study **(MammaRec)**												
	Observational (cohort) - non-randomized matched-pair study **(MammaLym)**												
	Observational (cohort) - non-randomized matched-pair study **(MammaMet)**												

Figure 2. Funnel plot (not considered for <10 studies)

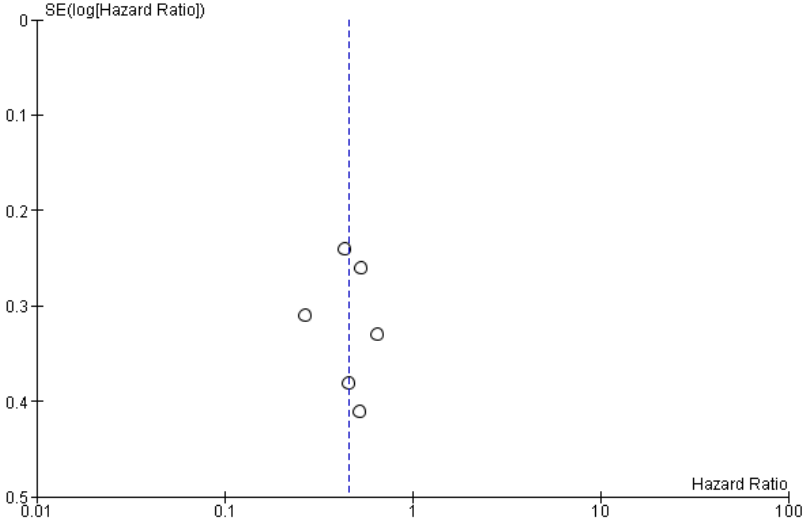

Table 4. Evidence profile table for survival

Quality Assessment							Summary of findings						
							Nr. of Patients		Effect			Quality grading	
Nr. of Studies	Design	Limitations	Inconsistency	Indirectness	Imprecision	Publication bias	Mistletoe	Control	Relative	Absolute			
									Hazard Ratio 95% CI	Control Risk	Risk Difference 95% CI	Start grade	Final grade
6	Obs	Serious limitations[1]	No serious inconsistency	No serious indirectness	No serious imprecision[2]	Unlikely	1029	1051	0.46 (0.36 - 0.58)	Not estimable[3]	Not estimable[3]	LOW	VERY LOW[4]

1. One study with no serious limitations, but 5 lacking accuracy and precision of the data, missing data on treatment regimes, some prognostic factors not included
2. Estimation of OIS not exact, but even worst estimation still met OIS
3. Calculations not possible with data disclosed within articles, limitation of not asking authors for needed missing data
4. Rated down for limitations

Appendix 3. Standardized quality assessment tables – TUMOUR RESPONSE

Table 1. Quality assessment for study limitations (risk of bias)

Observational												
Study	**Design**	**Limitations (risk of bias / internal validity)**										
		Failure to develop and apply appropriate eligibility criteria (inclusion of control population)	Under- or overmatching in case-control studies	Selection of exposed and unexposed in cohort studies from different populations	Flawed measurement of both exposure and outcome	Differences in measurement of exposure	Differential surveillance for outcome in exposed and unexposed in cohort studies	Failure to adequately control confounding	Failure of accurate measurement of all known important prognostic factors	Failure to match for prognostic factors and/or lack of adjustment in statistical analysis	Incomplete follow-up	Result
Grossarth-Maticek 2004 A	Observational (cohort) - prospective randomized matched-pair study **(MammaRand)**	Inclusion and exclusion criteria well developed and applied (pg 287), but due to study commencement in 1968, no written protocol, no pre-specified formulation of statistical hypotheses, no sample-size calculations in advance.	no	no	Even though predetermined case report forms were used, checked for consistency, missing data, and independent reviews on overall quality were performed (pg 288), authors recognize that accuracy and precision of the data are poor (exact dates of first diagnosis, operation, some data assessments and matching not available). Since this applies to both groups it would not affect one more than the other, but affects the overall quality of the results (not accurate).	no - predetermined case report forms, checked for consistency, missing data, ondependent reviews on overall quality (pg 288)	no	No data on doses, variation in dose, breaks in therapy, host trees etc (mistletoe therapy was prescribed by attending doctors, not study physicians)	Some medical prognostic factors were either not recorded throughout all cases, or not recorded at all (i.e. steroid receptor; histopathological type).	no -adequate match for the considered prognostic factors (tables 3 and 4 of supplemental material)	21 patients (6 from the treatment group) dropped out of the study (27%), so 21 matched-pairs were eliminated (from the 38). Does not affect randomization (performed pairwise and both partners were eliminated), but very small size. Patients were included in the final analysis (ITT).	Serious limitations
	Observational (cohort) - non-randomized matched-pair study **(Mamma)**										8% attrition bias. Impairments of internal validity due to drop-outs neutralized by exclusion of the corresponding match. All patients included in the complete set analyses.	

Table 2. Quality assessment for inconsistency and imprecision (first component: local recurrences)

Local recurrences									
Study	**Design**	**Inconsistency** - In relative, not absolute, measures of effect (RR/HR/OR) - Rate down if:					**Imprecision (around absolute, rather than relative effects)**		
		Point estimates vary widely across studies	Confidence intervals (Cis) show minimal or no overlap	Statistical test for heterogeneity shows a low p-value (<0.10 suggested by Cochrane for few	I^2 is large	**Result**	If the OIS (optimal information size) is not met, rate down for imprecision unless sample size very large.	95% Confidence Interval excludes no effect (CI around RR excludes 1.0)	**Result** If OIS met and 95% CI around effect excludes 1.0, no need to rate down for imprecision.
Grossarth-Maticek 2006 A	Observational (cohort) - prospective randomized matched-pair study **(MammaRand)** Observational (cohort) - non-randomized matched-pair study **(Mamma)**	Point estimates do not vary widely across studies	Confidence intervals (CIs) show reasonable overlap	0.9419	I^2 is zero (negative values are put equal to zero)	No serious inconsistency	OIS not met	Pooled HR excludes 1.0	Serious imprecision

Local recurrences								
	ln(HR)	SE(ln(HR))	wi = 1/VAR	Q	Δ2	wi*		I-square
MammaRand	-0.82	0.59	2.87	0.0053123672	0.00000	2.87		-18724%
Mamma	-0.87	0.35	8.16			8.16		Q: Chi test 0.94190
				Pes	random effects			
				-0.85698	PeDL	SE of PeDL		
					-0.85698	0.30102	95% CI	
				HR	0.42		0.24	0.77

Figure 1. Forest plot

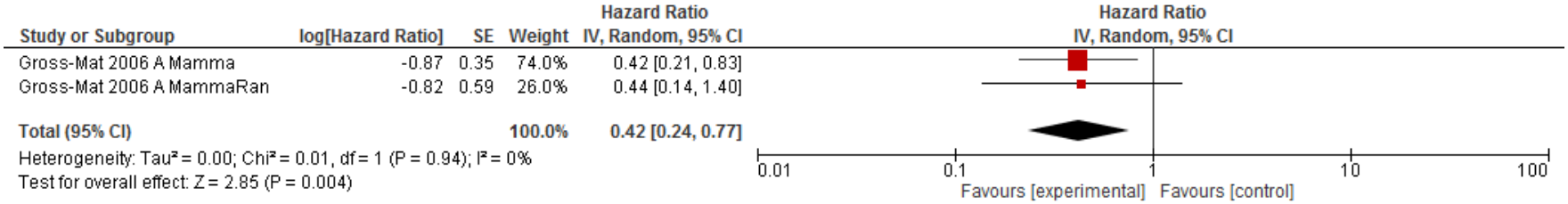

Table 3. Quality assessment for inconsistency and imprecision (second component: distant metastases)

Distant metastases									
Study	**Design**	**Inconsistency** - In relative, not absolute, measures of effect (RR/HR/OR) - Rate down if:					**Imprecision (around absolute, rather than relative effects)**		
		Point estimates vary widely across studies	Confidence intervals (Cis) show minimal or no overlap	Statistical test for heterogeneity shows a low p-value (<0.10 suggested by Cochrane for few	I^2 is large	**Result**	If the OIS (optimal information size) is not met, rate down for imprecision unless sample size very large.	95% Confidence Interval excludes no effect (CI around RR excludes 1.0)	**Result** If OIS met and 95% CI around effect excludes 1.0, no need to rate down for imprecision.
Grossarth-Maticek 2006 A	Observational (cohort) - prospective randomized matched-pair study **(MammaRand)** Observational (cohort) - non-randomized matched-pair study **(Mamma)**	Point estimates do not vary widely across studies	Confidence intervals (CIs) show reasonable overlap	0.477	I^2 is zero (negative values are put equal to zero)	No serious inconsistency	OIS not met	Pooled HR excludes 1.0	Serious imprecision

Distant metastases								
	ln(HR)	SE(ln(HR))	wi = 1/VAR	Q	Δ2	wi*		I-square
MammaRand	-0.69	0.37	7.30	0.5058	0.00000000	7.30		-98%
				Q: Chi test				
Mamma	-1.02	0.28	12.76	0.48		12.76		
				Pes	random effects			
				-0.90	PeDL	SE of PeDL		
					-0.90	0.22	95% CI	
				HR	0.41		0.26	0.63

Figure 2. Forest plot

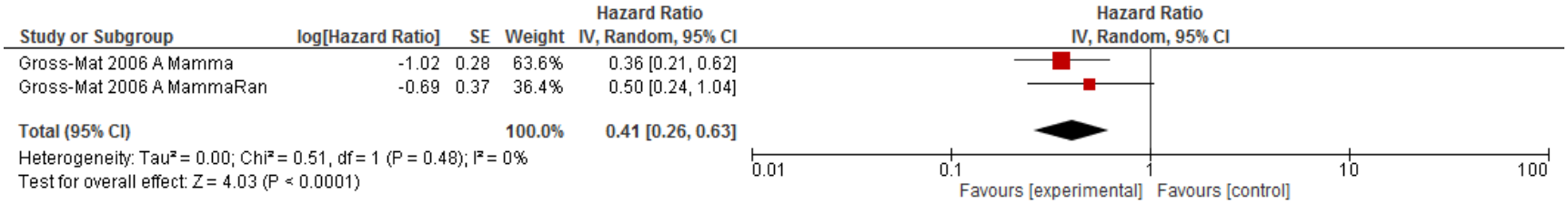

Table 4. Quality assessment for indirectness and publication bias

Study	Design	Indirectness						Publication Bias					
		Differences in population (applicability)	Differences in interventions (applicability)	Differences in outcome measures (surrogate outcomes)	Indirect comparisons because of biased head-to-head comparisons (e.g. industry)	Mechanism of action	Result	Early positive studies, small in size, preliminary and pilot studies	Non-English speaking countries might submit negative studies to local journals, "gray literature"	Double counting	Industry sponsored (or likely to be) or conflict of interest	Assymetrical funnel plot	Result
Grossarth-Maticek 2004 A	Observational (cohort) - prospective randomized matched-pair study **(MammaRand)** Observational (cohort) - non-randomized matched-pair study **(Mamma)**	not the case	not the case	not the case	not the case	not the case	No serious indirectness	Unlikely	Possible	no	not the case	Not considered (too few studies)	Unlikely

Table 5. Evidence profile table for tumour response

Quality Assessment							Summary of findings						
							Nr. of Patients		Effect			Quality grading	
Nr. of Studies	Design	Limitations	Inconsistency	Indirectness	Imprecision	Publication bias	Mistletoe	Control	Relative	Absolute			
									Relative Risk 95% CI	Control Risk	Risk Difference 95% CI	Start grade	Final grade
Local recurrences 2	Obs	Serious limitations[1]	No serious inconsistency	No serious indirectness	Serious imprecision[2]	Unlikely	122	122	0.42 (0.24 - 0.77)	Not estimable[3]	Not estimable[3]	LOW	VERY LOW[4]
Distant metastases 2	Obs						122	122	0.41 (0.23 - 0.66)				

1. Studies lacking accuracy and precision of the data, missing data on treatment regimes, some prognostic factors not included
2. Optimal Information Size not met
3. Calculations not possible with data disclosed within articles, limitation of not asking authors for needed missing data
4. Rated down for limitations and imprecision

Obs.: tumour response only partially assessable, for proportional hazard model adequacy not fulfilled for lymphatic metastases and all events pooled.

Appendix 4. Standardized quality assessment tables – QUALITY OF LIFE

Table 1. Quality assessment for study limitations (risk of bias)

RCT													
Study	**Design**	**Limitations (risk of bias / internal validity)**											
		Lack of allocation concealment	Lack of blinding	Incomplete accounting of patients and outcome events	Selective outcome reporting bias	Stopping early for benefit	Use of unvalidated outcome measures	Carryover effects in crossover trial	Recruitment bias in cluster-randomized trials	Baseline values	Unappropriate consideration of Intention-to-treat principle	Conclusion	Result
Tröger 2009	Prospective randomized open label clinical trial (pilot study)	Low risk. Probably done. Quote: "Allocation concealment was implemented by using sealed envelopes" (…)	Low risk. Even though blinding was probably not done: quote: " CRDT (Clinical Research Dr. Tröger) was responsible for planning, conduct, monitoring, and analysis of the study." , since the EORTC questionnaires were answered by the patients themselves this was not considered a high risk of bias in this case.	Low risk. One patient control group withdrawn for heart disease (3%); less than 1% missing data at both groups for both assessed parameters	Low risk. Planned outcomes reported.	Low risk.	Low risk (absent).	Low risk (absent).	Absent	Balanced (pg 38)	Absent (pg37: "statistical analysis was performed on the intention-to-treat population").	Low risk of bias for all key criteria	No serious limitations
Semiglazov 2006	RCT (double blind)	Low risk. Probably done. Quote: "Patients were allocated to the treatment groups on the basis of a computer-generated randomisation list."	Low risk. Double blind.	Low risk. 4 patients lost to placebo group = 2% (2 decision of the patient, 1 serious adverse effect, 2 patients moved). 1 patient lost to test group = 0,6% (decision of patient)	Low risk. Planned outcomes reported.	Low risk.	Low risk (absent).	Low risk (absent).	Absent	Balanced between treatment groups with respect to demographics and medical history, not with respect to QoL scale scores (stronger restrtiction at test group) and histological classification (lower rate of invasive tumors inplacebo group), evaluation complemented by adjusted analyses of variance. Considered low risk because is in favor of placebo group.	4% attrition at test and 5% at control group	Low risk of bias for all key criteria	
Semiglasov 2004	RCT (double blind)	Low risk. Probably done. Quote: "Patients were allocated to the treatment groups on the basis of a computer-generated randomisation list."	Low risk. Double blind.	Low risk. 4% (11 patients lost: 4 to adverse effects, 4 decision of patient, 3 other reasons) ITT analysis.	Low risk. Planned outcomes reported.	Low risk.	Low risk (absent).	Low risk (absent).	Absent	Balanced between treatment groups with respect to demographics and medical history, not with respect to QoL scale scores (but not statistically significant), so considered low risk.	3% attrition at test and 6% at control group	Low risk of bias for all key criteria	

Table 2. Quality assessment for inconsistency and imprecision

Study	Design	Inconsistency - In relative, not absolute, measures of effect (RR/HR/OR) - Rate					Imprecision (around absolute, rather than relative effects)		
		Point estimates vary widely across studies	Confidence intervals (Cis) show minimal or no overlap	Statistical test for heterogeneity shows a low p-value (<0.10 suggested by Cochrane for few	I^2 is large	**Result**	If the OIS (optimal information size) is not met, rate down for imprecision unless sample size very large.	95% Confidence Interval excludes no effect (CI around RR excludes 1.0)	**Result** If OIS met and 95% CI around effect excludes
Tröger 2009	Prospective randomized open label clinical trial (pilot study)	Not possible to perform calculations, since no common score was used and the differences between them do not allow a transformation of results.							
Semiglazov 2006	RCT (double blind)								
Semiglasov 2004	RCT (double blind)								

Table 3. Quality assessment for indirectness and publication bias

Study	Design	Indirectness						Publication Bias					
		Differences in population (applicability)	Differences in interventions (applicability)	Differences in outcome measures (surrogate outcomes)	Indirect comparisons because of biased head-to-head comparisons (e.g. industry)	Mechanism of action	Result	Early positive studies, small in size, preliminary and pilot studies	Non-English speaking countries might submit negative studies to local journals, "gray	Double counting	Industry sponsored (or likely to be) or conflict of interest	Assymetrical funnel plot	Result
Tröger 2009	Prospective randomized open label clinical trial (pilot study)	not the case	not the case	not the case	not the case	not the case	No serious indirectness	Small in size; Tröger and Semiglasov 2004 are pilot studies	Possible	not the case (revised for dupplicates)	Semiglazov 2006 and Semiglazov 2004 received funding/medication provision from Madaus GmbH, but all studies were performed by independent research centra	Not considered (too few studies)	Possible Publication bias
Semiglazov 2006	RCT (double blind)												
Semiglasov 2004	RCT (double blind)												

Table 4. Evidence profile table for tumour response

Quality Assessment							Summary of findings						
							Nr. of Patients		Effect			Quality grading	
Nr. of Studies	Design	Limitations	Inconsistency	Indirectness	Imprecision	Publication bias	Mistletoe	Control	Relative	Absolute			
									Relative Risk 95% CI	Control Risk	Risk Difference 95% CI	Start grade	Final grade
3	RCT	No serious limitations	Not assessable[1]	No serious indirectness	Not assessable[1]	Possible	271	277	Not assessable[1]	Not assessable[1]	Not assessable[1]	HIGH	Not assessable [2]

1. Not possible to perform calculations, since no common score was used and the differences between them do not allow a transformation of results.

2. Since assessments on inconsistency and imprecision were not possible, it is also not possible to know if the quality would have been downgraded or not.

Reference list

[1] National Center for Complementary and Alternative Medicine. Retrieved June 18th, 2012, from http://nccam.nih.gov/health/whatiscam

[2] Frass ME. Use and acceptance of complementary and alternative medicine among the general population and medical personnel: a systematic review. *The Ochsner Journal*, 2012 Spring; 12(1): 45-56.

[3] CAMbrella – pan-European research network for complementary and alternative medicine (CAM). European Policy Brief. 2012 Nov; 16 pages.

[4] National Center for Complementary and Alternative Medicine.(n.d.) Retrieved November 6th, 2012, from

http://nccam.nih.gov/news/camstats/costs/costdatafs.htm

[5] Thomas KJ, Nicholl JP, Coleman P. Use and expenditure on complementary medicine in England: a population based survey. *Complementary Therapies in Medicine*, 2001; 9 (1):2-11.

[6] Xue CC, Zhang AL, Lin V, Da Costa C, Story DF. Complementary and alternative medicine use in Australia: a national population-based survey. *The Journal of Alternative and Complementary Medicine*, 2007; 13(6):643-50.

[7] Molassiotis A, Fernadez-Ortega P, Pud D, Ozden G, et al. Use of complementary and alternative medicine in cancer patients: a European survey. *Annals of Oncology: official journal of the ESMO*, 2005 Apr;16(4):655-63.

[8] *Weiger WA, Smith M, Boon H et al.* Advising patients who seek complementary and alternative medical therapies for cancer. *Ann Intern Med* 2002*;*137*:*889*–903.*

[9] *Adams J, Sibbritt DW, Easthope G et al.* The profile of women who consult alternative health practitioners in Australia*. Med J Aust* 2003*;*179*:*297*–300.*

[10] *Chrystal K, Allan S, Forgeson G et al.* The use of complementary/alternative medicine by cancer patients in a New Zealand regional cancer treatment centre. *NZ Med J* 2003*;*116*:*U296*.*

[11] *Lee MM, Chang JS, Jacobs B et al*. Complementary and alternative medicine use among men with prostate cancer in 4 ethnic populations. *Am J Public Health* 2002*;*92*:*1606*–1609.*

[12] *Eisenberg DM, Davis RB, Ettner SL et al*. Trends in alternative medicine use in the United States, 1990–1997: results of a follow-up national survey. *JAMA* 1998*;*280*:*1569*–1575.*

[13] Baars E. Evidence-based curative health promotion . Wageningen: Wageningen University, 2011.

[14] Witt C, Linde K. Clinical Research in Complementary and Integrative Medicine*.* Munich: Elsevier, 2011.

[15] Fønnebø V, Grimsgaard S, Walach H, Ritenbaugh C, Norheim AJ, MacPherson H et al. Researching complementary and alternative treatments – the gatekeepers are not at home. *BMC Medical Research Methodology* 2007; 7:7.

[16] Kienle GS, Kiene H, Albonico HU. Anthroposophic Medicine – Effectiveness, utility, costs, safety. Stuttgart: Schattauer; 2006: 213-214.

[17] Kienle GS, Glockmann A, Grugel R, Hamre HJ, Kiene H. Klinische Forschung zur anthroposophischen Medizin – Update eines "Health Technology Assessments"-Berichts und Status Quo. *Forschende Komplementärmedizin* 2011 October; 18:269-282.

[18] Kienle GS, Kiene H. Influence of Viscum album L (European mistletoe) extracts on quality of life in cancer patients: a systematic review of controlled clinical studies. Integrative Cancer Therapies 2010; 9(2): 142-157.

[19] Büssing, A. and (ed.), Mistletoe.*The Genus Viscum.* pp. 1-265, Hardwood Academic Publishers, Amsterdam (2000).

[20] Moraes WA, The epistemological foundations of medicine amplified by anthroposophy (As bases epistemológicas da medicina ampliada pela antroposofia). São Paulo, Associação Brasileira de Medicina Antroposófica, 2005.

[21] Medical Section of the School of Spiritual Science, International Federation of Anthroposophic Medical Associations (IVAA) and Association of Anthroposophic Physicians in Germany (GAÄD). Vademecum of Anthroposophic Medicines. *Der Merkurstab Journal of Anthroposophic Medicine*, 2009; 62 Suppl.

[22] Kienle, G. S. and H. Kiene, *Die Mistel in der Onkologie - Fakten und konzeptionelle Grundlagen.* Schattauer Verlag, Stuttgart, New York (2003).

[23] Kienle GS, Glockmann A, Schink M, Kiene H. Viscum album extracts in breast and gynaecological cancers: a systematic review of clinical and preclinical research. Journal of Experimental and Clinical Research, 2009 June; 28:79.

[24] Ostermann T, Raak C, Büssing A. Survival of cancer patients treated with mistletoe extract (Iscador): a systematic literature review. *BMC Cancer* 2009 Dec 18; 9:451.

[25] Büssing A, Raak C, Ostermann T. Quality of life and related dimensions in cancer patients treated with mistletoe extract (Iscador): a meta-analysis. *Evidence-Based Complementary and Alternative Medicine*. 2012: 219402 (8 pages).

[26] The World Health Organization - Worldwide Breast Cancer. Retrieved November 7th, 2012, from http://www.worldwidebreastcancer.com/map/

[27] *National Cancer Institute -Survellance Epidemiology and End Results* Retrieved November 7th, 2012, from http://seer.cancer.gov/statfacts/html/breast.html

[28] International Agency for Research on Cancer - IARC - Globocan. Retrieved November 7th, 2012, from http://globocan.iarc.fr/factsheets/cancers/breast.asp

[29] National Cancer Institute - A Snapshot of Breast Cancer. Retrieved November 7th, 2012, from http://www.cancer.gov/aboutnci/servingpeople/snapshots/breast.pdf

[30] Wanchai A, Armer J, Stewart R. Complementary and Alternative Medicine Use Among Women With Breast Cancer: A Systematic Review. *Clinical Journal of oncology nursing.* 2010 Aug;14(4):E45-55.

[31] Schöffski O, Emmert M. Comparative cost analyses of mistletoe preparations for breast carcinoma aftercare. [Vergleichende kostenanalyse der Nachsorge beim Mammakarzinom unter besonderer Berücksichtigung von Mistelpräparaten] *Deutsche Zeitschrift für Onkologie*, 2009; 41(3):119-125.

[32] CAMbrella – European research network for complementary and alternative medicine. Retrieved June 18th, 2012, from

http://www.cambrella.eu/home.php

[33] ECHAMP - European Coalition on Homeopathic and Anthroposophic Medicinal Products. Retrieved June 30th, 2013, from

http://www.echamp.eu/news/newsletter/newsletter-archive/2012/march/an-interview-with-dr-susanne-schunder-tatzber.html

[34] ECHAMP - European Coalition on Homeopathic and Anthroposophic Medicinal Products. Retrieved June 30th, 2013, from

http://www.echamp.eu/news/newsletter/newsletter-archive/2012/june/cam-events-in-brussels.html

[35] EPHA – European Publica Health Alliance. Retrieved June 30th, 2013, from http://www.epha.org/

[36] Cochrane CAM Field. Retrieved 30th June, 2013, from http://www.compmed.umm.edu/cochrane_about.asp

[37] Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, Alderson P et al. GRADE guidelines: 2 Framing the question and deciding on important outcomes. *Journal of Clinical Epidemiology,* 2011 April;64(4):395-400.

[38] Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: A new series of articles in the Journal of Clinical Epidemiology. *Journal of Clinical Epidemiology* 2011 April; 64: 380-382.

[39] Guyatt GH, Oxman A, Akl EA, Kunz R, Vist G, Brozek J, Norris S et al. GRADE guidelines 1: Introduction – GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology 2011*; 64: 383-394.

[40] The Grading of Recommendations Assessment, Development and Evaluation (short GRADE) Working Group. Retrieved June 5th, 2013, from http://www.gradeworkinggroup.org/society/ retrieved 5th June 2013

[41] Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *British Medical Journal,* 1996;312:71

[42] Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, Atkins D et al. GRADE guidelines: 9. Rating up the quality of evidence. *Journal of Clinical Epidemiology* 2011 December; 64(12): 1311-1316.

[43] Brozek JL, Bousquet J, Baena-Cagnani CE, Bonini S, Canonica GW et al.(2010) "Allergic Rhinitis and its Impact on Asthma (ARIA) guidelines: 2010 Revision. Journal of Allergy and Clinical Immunology, Volume 126, nr 3, pp. 467-476.)

[44] Balshem H, Helfand M, Schünemman HJ, Oxman A, Kunz R, Brozek J, Vist GE et al. GRADE guidelines: 3. Rating the quality of the evidence. *Journal of Clinical Epidemiology* 2011 April; 64(4): 401-406.

[45] Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V et al. GRADE guidelines: 4. Rating the quality of evidence – study limitations (risk of bias). *Journal of Clinical Epidemiology* 2011 April; 64(4): 407-415.

[46] Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, Deveraux PJ et al. GRADE guidelines: 6. Rating the quality of evidence – imprecision. *Journal of Clinical Epidemiology* 2011 December; 64(12): 1283-1293.

[47] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P et al. GRADE guidelines: 7. Rating the quality of evidence – inconsistency. *Journal of Clinical Epidemiology* 2011 December; 64(12): 1294-1302.

[48] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P et al. GRADE guidelines: 8. Rating the quality of evidence – indirectness. *Journal of Clinical Epidemiology* 2011 December; 64(12): 1303-1310.

[49] Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, Alonso-Coello P et al. GRADE guidelines: 5. Rating the quality of evidence – indirectness. *Journal of Clinical Epidemiology* 2011 December; 64(12): 1277-1282.

[50] Guyatt GH, Oxman AD, Santesso N, Helfand M, Vist G, Kunz R, Brozek J et al. GRADE guidelines: 12. Preparing Summary of Findings tables – binary outcomes. *Journal of Clinical Epidemiology* 2013 Feb;66(2):158-72.

[51] Guyatt GH, Oxman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P, Atkins D et al. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. *Journal of Clinical Epidemiology* 2013 Feb;66(2):151-157.

[52] Son GS, Ryu WS, Kim HY, Woo SU, Park KH, Bae JW. Immunologic response to mistletoe extract (*Viscum album L.)* after conventional treatment in patients with operable breast cancer. *Journal of Breast Cancer*, 2010 March; 13(1):14-8.

[53] Tröger W, Jezdić S, Ždrale Z, Tišma N, Hamre H. Quality of life and neutropenia in patients with early stage breast cancer: a randomized pilot study comparing additional treatment with mistletoe extract to chemotherapy alone. *Breast Cancer: Basic and Clinical Research*, 2009 Jul 6;3:35-45.

[54] Beuth J, Schneider B, Schierholz JM. Impact of complementary treatment of breast cancer patients with standardized mistletoe extract during aftercare: a controlled multicenter comparative epidemiological cohort study. *Anticancer Research*, 2008 Jan-Feb; 28(1B):523-7.

[55] Grossarth-Maticek R, Ziegler R. Prospective controlled cohort studies on long-term therapy of breast cancer patients with a mistletoe preparation (Iscador). Forschende Komplementärmedizin 2006 Oct;13(5):285-92.

[56] Grossarth-Maticek R, Ziegler R. Randomised and non-randomised prospective controlled cohort studies in matched-pair design for the long-term therapy of breast cancer patients with a mistletoe preparation (Iscador): a re-analysis. European Journal of Medical Research, 2006 Nov 30;11(11):485-95.

[57] Semiglasov VF, Stepula VV, Dudov A, Schnitker J, Mengs U. Quality of life is improved in breast cancer patients by standardized mistletoe extract PS76A2 during chemotherapy and follow-up: a randomized, placebo-controlled, double-blind, multicenter clinical trial. Anticancer Research 2006 Mar-Apr;26(2B):1519-29.

[58] Büssing A, Bischof M, Hatzmann W, Bartsch F, Soto-Vera D, Fronk EM, Gmeindl M et al. Prevention of surgery-induced suppression of granulocyte function by intravenous application of a fermented extract from *Viscum album L.* in breast cancer patients. Anticancer Research 2005 Nov-Dec;25(6C):4753-7.

[59] Semiglasov VF, Stepula VV, Dudov A, Lehmacher W, Mengs U. The standardized mistletoe extract PS76A2 improves QoL in patients with breast cancer receiving adjuvant CMF chemotherapy: a randomized, placebo-controlled, double-blind, multicenter clinical trial. Anticancer Research 2004 Mar-Apr;24(2C):1293-302.

[60] Bock PR, Friedel WE, Hanisch J, Karasmann M, Schneider B. Retrolective, comparative, epidemiological cohort study with parallel groups design for evaluation of efficacy and safety of drugs with 'well-established use'. *Forschende Komplementärmedizin* 2004; 11(Suppl 1): 23-29.

[61] Brozec JL, Baena-Cagnani CE, Bonini S, Canonica GW, Rasi G, van Wijk RG, Zuberbier T et al. Methodology for development of the Allergic Rhinitis and its impact on Asthma Guideline 2008 update. Allergy, 2008 Jan;63(1):38-46.

[62] Swiglo BA, Murad MH, Schünemann HJ, Kunz R, Vigersky RA, Guyatt GH, Montori V. A case for clarity, consistency, and helpfulness: state-of-the-art clinical practice guidelines in endocrinology using the grading of recommendations, assessment, development, and evaluation system. *The Journal of Clinical Endocrinology and Metabolism*, 2008 March; 93(3):666-673.

[63] Thornton J, Alderson P, Tan T, Turner C, Latchem S, Shaw E, Ruiz F et al. Introducing GRADE across the NICE clinical guideline program. Journal of Clinical Epidemiology 2013; 66:124-131.

[64] Duclos P, Durrheim DN, Reingold AL, Bhutta ZA, Vannice K, Rees H. Developing evidence-based immunization recommendations and GRADE. *Vaccine* 2012 December; 31 (1):12-19.

[65] Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. *British Medical Journal*. 2006 September; 333 (7568): 597-600.

[66] Chawla SP, Buzdar AU, Hortobagyi GN, Blumenschein GR. Tumor-associated fever in breast cancer. *Cancer*. 1984 Apr 1;53(7): 1596-9.

[67] Williams PD, Ducey KA, Sears AM, Williams AR, Tobin-Rumelhart SE, Bunde P. Treatment type and symptom severity among oncology patients by self-report. *International Journal of Nursing Studies*, 2001; 38(3):359-367.

[68] Breastcancer.org, USA. Retrieved May 28th, 2013, from http://www.breastcancer.org/treatment/side_effects

[69] American Cancer Society. Retrieved May 28th, 2013, from http://www.cancer.org/cancer/breastcancer/index

[70] Dickman A, Schneider J. The Syringe Driver. Oxford: Oxford University Press; 2011.

[71] Frohman EM. Dr. Elliot Frohman on managing injection site reactions. *Inside MS*, 2005 Apr-May:6263.

[72] Auerbach, L., V. Dostal, I. Václavik-Fleck, E. Kubista, A. Rosenberger, S. Rieger, W. Tröger and J. M. Schierholz, Signifikant höherer Anteil aktivierter NK-Zellen durch additive Misteltherapie bei chemotherapierten Mamma-Ca-Patientinnen in einer prospektiven randomisierten doppelblinden Studie. In.Fortschritte in der Misteltherapie. Aktueller Stand der Forschung und klinischen Anwendung. (Ed. R. Scheer et al.) pp. 543-554, KCV Verlag, Essen (2005).

[73] Rostock, M. and R. Huber. Randomized and double-blind studies - demands and reality as demonstrated by two examples of mistletoe research. *Forschende Komplementärmedizin,* 2004 Aug; 11 Suppl 1:18-22.

[74] Ioannidis JP, Evans SJ, Gotzsche PC, O'Neill RT, Altman DG, Schulz K, Moher D. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Annals of Internal Medicine 2004; 141(10):781-788

[75] Common Terminology Criteria for Adverse Events (CTCAE). National Cancer Institute – Cancer Therapy Evaluation Program. Retrieved May 28th, 2013, from http://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm.

[76] MSSO- Maintenance and Support Services Organization. MedDRA - the Medical Dictionary for Regulatory Activities. Regulatory information. Retrieved May 28th, 2013, from http://www.meddramsso.com/public_aboutMedDRA_regulatory.asp

[77] European Comission - Detailed guidance on the collection, verification and presentation of adverse event/reaction reports arising from clinical trials on medicinal products for human use ('CT-3'). *Official Journal of the European Union*, 2011 June 11, C172/2. Retrieved May 28th, 2013, from http://ec.europa.eu/health/files/eudralex/vol-10/2011_c172_01/2011_c172_01_en.pdf

[78] Tröger W, Ždrale Z, Stanković N, Matijašević M. Five-year follow-up of patients with early randomized stage breast cancer after a study comparing additional treatment with Viscum album (L.) extract to chemotherapy alone. *Breast Cancer: Basic and Clinical Research* 2012;6:173-80

[79] Rehfuess E, Akl E. Current experience with applying GRADE approach to public health interventions: an empirical study. *BMC Public Health*, 2013 January; 13:9.