Université d'Ottawa · University of Ottawa

THE ASSOCIATION BETWEEN ESTROGEN-PROGESTIN REPLACEMENT THERAPY AND THE RISK OF BREAST CANCER AMONG POST-MENOPAUSAL WOMEN: A SYSTEMATIC REVIEW AND META-ANALYSIS

by

LYNN MYER BRODSKY

Thesis submitted to
the School of Graduate Studies and Research
in partial fulfilment of the requirements for the
MSc degree in Epidemiology

University of Ottawa

JANUARY 2002

Abstract

Postmenopausal women rely on hormone replacement therapy to manage menopausal symptoms. Few studies have assessed the consequences of estrogen progestin replacement therapy (EPRT), a formulation prescribed to postmenopausal women with an intact uterus. The aim of this thesis was to conduct a systematic review and meta-analysis assessing the association between breast cancer risk and EPRT use among postmenopausal women.

Electronic databases, handsearching and advise from experts were used to locate studies. A fixed effects model was used to calculate the combined estimate and its standard error. Statistical tests for association and homogeneity were performed.

The association between breast cancer risk and EPRT exposure was significant for both case control (OR = 1.38; 95% CI 1.23-1.56; $\chi^2 = 48.30$, $p < 0.0001$) and cohort studies (RR= 1.42; 95% CI 1.29-1.57; $\chi^2 = 27.95$, $p < 0.0001$). Results indicate that postmenopausal women on EPRT have a higher breast cancer risk, particularly with long term use and testosterone derived formulations.

Abstract

Objective: To conduct a systematic review and meta-analysis of the pertinent literature assessing the association between exposure to estrogen progestin replacement therapy (EPRT) and the risk of breast cancer among postmenopausal women.

Data Sources: The electronic data bases Pre-Medline, Medline, Embase, Cancerlit, Current Contents, Dissertation Abstracts and Biological Abstracts as well as Pubmed, and Biomedcentral. The internet search engine Google was also used. Hand searching of key journals and conference proceedings was also performed. Advice from experts was also sought and pertinent references from articles identified from the electronic search were retrieved.

Review Methods:
Selection criteria—Case control and cohort studies (design) that examined the association between exposure to EPRT (intervention) and breast cancer (outcome) among postmenopausal women (population) were selected. Studies were excluded if i) exposure to estrogen and progestin was restricted to oral contraceptive use and ii) outcomes were intermediate or reported on tumour type or mortality due to breast cancer. If the same subject was reported in multiple publications, only the most recent analysis was used and subsequent reports were excluded.

Method for validity assessment—The Newcastle-Ottawa Scale was used to assess the quality of nonrandomized studies (both case controls and cohorts) examined in this investigation.

Data abstraction—Two reviewers were involved in selecting the studies for inclusion in the meta-analysis and assessing study quality. All other aspects of the review were completed independently. Standardized forms were used to extract the data from each study. Data on study design, location, methodology, exposure and outcome assessment, characteristics of the population and controlled risk factors were collected.

Data synthesis—A fixed effects model was used to calculate the combined estimate and its standard error for case control (odds ratio) and cohort studies (relative risk). Statistical tests for association (chi-square test) and homogeneity (Cochran's Q test) were performed.

Results: Eleven case control and four cohort studies were included in the meta-analysis. The combined odds ratio estimate was 1.42 (95% CI 1.29, 1.57) and the combined relative risk was 1.38 (95% CI 1.23, 1.56). The association between exposure to EPRT and breast cancer was significant for both case control studies ($\chi^2 = 48.30$, $p < 0.0001$) and cohort studies ($\chi^2 = 27.95$, $p < 0.0001$), suggesting a higher risk of breast cancer for women exposed to EPRT. There was no significant heterogeneity within the case control studies ($\chi^2 = 12.32$, $p = 0.27$), or within the cohort studies ($\chi^2 = 2.80$, $p = 0.42$). There was variation between studies. Subgroup and sensitivity analyses were completed in an attempt to identify factors contributing to this variation. Variation was explained, in part,

by differences in study quality (higher quality studies reported higher risk), duration of EPRT exposure (longer exposure duration was associated with higher risk), study location (European studies reported higher risk than North American studies), recency of EPRT use and the chemical nature of the progestin formulation (use of testosterone derived progestins were associated with higher risk than progesterone derived progestins). The sensitivity analyses revealed that the inclusion and exclusion of rejected studies did not influence the summary estimates. Results from the chronologically cumulative meta-analysis revealed that over time, the breast cancer risk associated with EPRT exposure has increased. As study quality declined, the results from the cumulative meta-analysis indicated that the reported breast cancer risk also declined.

Conclusions: Postmenopausal women on EPRT experience a higher risk of developing breast cancer compared to women that have never taken any hormone replacement therapy. This was particularly evident with long term (greater than 10 years) EPRT use and formulations prepared with testosterone derived progestins.

Acknowledgements

I would like to thank my supervisor, Dr. George Wells for the guidance and support he provided throughout this project. I am most appreciative of all the time and effort he devoted to our regular meetings, readily explaining the intricacies of performing meta-analyses on nonrandomized studies and answering all of my analysis questions. I would also like to thank my advisor, Dr. Eva Grunfeld for her exceptional advice and thorough reviews of my thesis drafts. She was particularly helpful in the preparation of this thesis and explaining relevant issues pertaining to breast cancer and hormone replacement therapy. I would also like to thank my thesis examiners, Dr. Don Wigle and Dr. Ineke Neutel for their comments.

Dr. Bob Spasoff, Fay Draper, Mariella Peca and Anne Gray provided advice and support of various kinds throughout my program. I would like to thank them for their kindness and understanding.

I would also like to recognise the technical assistance provided by various information specialists who helped me develop my search strategies. Jesse McGowan and Margaret Sampson provided excellent support. Joan Peterson also provided excellent assistance with the screening of articles and evaluation of study quality. The staff at the RGN, Claire and Normand must also be thanked for their help with interlibrary loans.

I would also like to thank Dr. Anne Cranney, Jacq Dinnes, Dr. Emily White, Victoria Kirsh, Dr. Chi-Ling Chen and Margaret Sampson for their generosity in providing me with detailed information about their research.

The completion of this thesis would not have been possible without the encouragement and continued support from my husband, Michael.

Table of contents

List of tables

List of figures

.

1. INTRODUCTION

1.1 Menopause

There are over 4.7 million women in Canada over the age of fifty[1], the average age when menopause occurs.[2] Menopause signals the end of a woman's reproductive period, when her levels of estrogen and progesterone decline and her menstrual cycle stops. Since the life expectancy of the average Canadian woman is estimated to be 81 years of age[3], she can expect to be postmenopausal for more than one third of her life.

Since the 1930's, many physicians have advocated the use of hormone replacement therapy (HRT) to compensate for declining levels of hormones.[4] Early formulations of HRT were estrogen based and were referred to as estrogen replacement therapy (ERT). ERT has been prescribed to postmenopausal women to manage the short and long term consequences of menopause.[5;6] The short term consequences include hot flashes and night sweats (vasomotor symptoms) and vaginal dryness and urinary problems (genitourinary symptoms).[7] Longer term consequences of menopause include accelerated bone loss and more rapid progression of cardiovascular disease.[8;9]

1.2 Benefits of estrogen replacement therapy

ERT use has provided two major health benefits for postmenopausal women. ERT has been found to be effective in controlling the vasomotor and genitourinary symptoms of menopause.[10;11] As well, Levin[12] states that ERT is the "most effective single modality for preventing osteoporosis in post menopausal women". It has been found to help maintain bone density and thus help reverse bone loss.[13] Results from the Postmenopausal Estrogen/Progestin Interventions (PEPI) Trial[14] reported increased bone-

mineral density among postmenopausal women on ERT compared to postmenopausal women on placebo.

Until recently, it was also believed that ERT afforded users protection against cardiovascular disease.[15-17] However the findings from over 40 nonrandomized studies have not been confirmed by the results from randomized controlled trials (RCTs).[18] The results from the Heart and Estrogen/progestin Replacement Study (HERS) suggest that there are no cardiovascular benefits associated with HRT use.[19]

Several studies have also suggested that HRT may reduce the risk of colorectal cancer[20] and Alzheimer's Disease[21]. However, evidence from other studies refute these findings.[22;23]

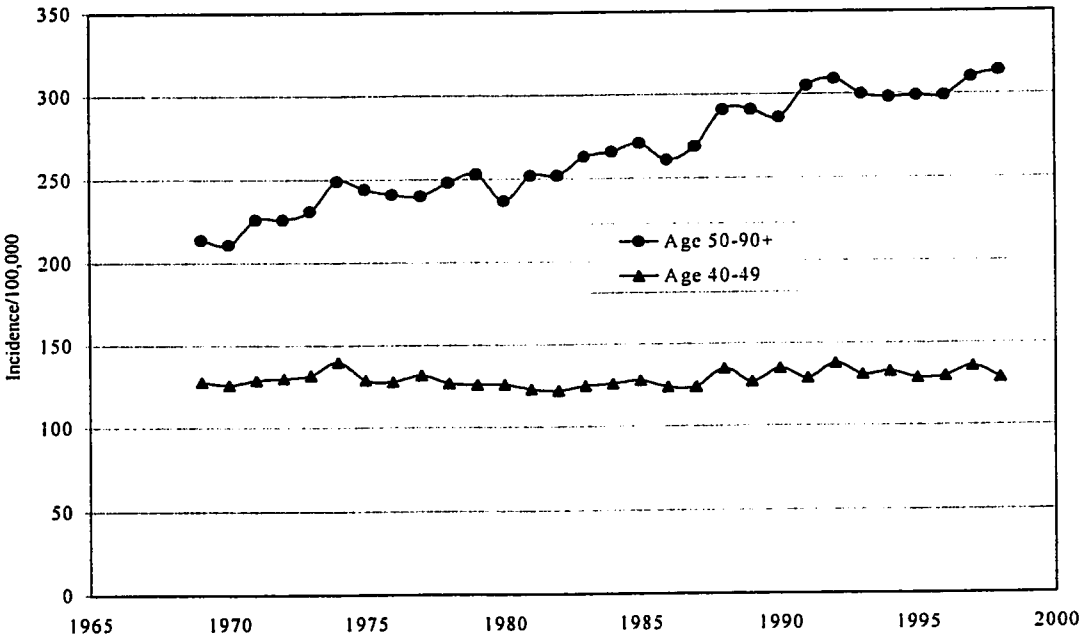
1.3 Risks of estrogen replacement therapy

In order to be effective at managing osteoporosis, ERT must be taken for at least five years and continued benefit requires continued use.[24] Cauley et al[25] report that ten years after women stop taking HRT, their bone density and fracture risk is similar to women who had never been on HRT. Given the potential for long term exposure to HRT, it is essential that there is a clear understanding of its health consequences.

There have been over fifty nonrandomized studies that have examined the association between ERT and breast cancer[8], along with six different meta-analyses[26-31]. All have essentially reached the same conclusion: there is no increased risk associated with short term use of ERT. Studies that have looked at long term use (more than five years) reported a higher risk[28;30;31].

Breast cancer is of particular concern to postmenopausal Canadian women as it has the highest incidence of all cancers in women.[32] It is estimated that there will be 15,100 new cases of breast cancer this year among postmenopausal women alone, representing 77% (n = 19,500) of all breast cancer cases in 2001.[32] Over 4,700 postmenopausal women are expected to die from breast cancer this year, representing 85% (n = 5,500) of all breast cancer mortality. The age-standardized incidence rates (ASIR; adjusted to the age distribution of the 1991 Canadian population) of breast cancer among postmenopausal women have been increasing steadily since 1970 (Figure 1.1). The rate among women 50 years of age and over has increased from 211/100,000 in 1970 to 310/100,000 in 1997—a 47% increase.[33]

Figure 1.1 Incidence of breast cancer among pre- and postmenopausal Canadian women, 1969-1998



While the breast cancer risk associated with short term use of ERT is reportedly low, Grady et al[34] found that the risk of endometrial cancer associated with ERT use is quite high. In a meta-analysis that examined the results of 35 nonrandomized studies, they found that the risk of endometrial cancer was anywhere from two to twenty times higher in women on ERT. The PEPI Trial revealed that adding progestins to the ERT significantly reduced the risk of endometrial cancer, such that women on estrogen progestin replacement therapy (EPRT) experienced a lower risk. As a result, EPRT rather than ERT is increasingly being prescribed to women with an intact uterus to treat menopausal symptoms. It is still not clear what effect these added progestins have on the risk of breast cancer.

Early studies examining the effects of EPRT on breast cancer [35-37] found no increased risk of breast cancer among women on EPRT. However, the results of more recent studies [38-40] suggest an increased risk of breast cancer associated with EPRT use. To date there has been no systematic review of the literature that incorporates the results of all of these studies, particularly those published since 1995 or one that focuses exclusively on the association between EPRT and breast cancer.

1.4 Objective

The aim of this study was to examine the association between exposure to EPRT and breast cancer among postmenopausal women. This was achieved by means of a systematic review of the literature and, where appropriate, a meta-analysis.

2. BACKGROUND

In this chapter, a description of menopause and its short and long term physiological consequences will be presented. The risks and benefits associated with ERT use, prescribed to manage these menopausal symptoms, will also be discussed. The rationale behind the development of EPRT will be highlighted and the need for a systematic review to assess the association between breast cancer and EPRT will be examined. An overview of the current meta-analytic methodologies for nonrandomized studies will be provided, with particular emphasis on the process of quality assessment and one assessment tool designed specifically for this purpose.

2.1 Menopause: definition and description

During her reproductive years, a woman produces a series of five hormones that regulate her monthly ovulation/menstrual cycle. The release of these hormones is coordinated through negative feedback. Gonadotropin-releasing hormone (GnRH) is produced in and secreted by the hypothalamus in response to low serum levels of estrogen and progesterone, two hormones produced in the ovaries. At the start of the ovulation and menstrual cycle, GnRH stimulates the pituitary to release two gonadotropin hormones: follicle stimulating hormone (FSH) and lutenizing hormone (LH). FSH stimulates the growth and development of a follicle in the ovary and during its development, estrogen is released. Increasing levels of estrogen stimulate the release of LH, which induces the final maturation of the follicle. LH also leads to the formation of the corpus luteum, which secretes more estrogen and progesterone. The now high levels

of estrogen and progesterone inhibit further release of GnRH and promote development of the endometrium (the inner lining of the uterus). As levels of FSH and LH subsequently decline, the corpus luteum and the endometrium break down. This leads to a drop in estrogen and progesterone levels and the shedding of the endometrium as menstrual fluid. The drop in hormone levels leads to the release of GnRH, reinitiating the two cycles.[41]

These cycles begin when a young woman enters puberty (age 12-18) and end 30-40 years later at menopause. Menopause includes the cessation of the menstrual cycle as well as changes in serum levels of the five hormones discussed above. Several years before menopause begins, estrogen and progesterone levels decline and GnRH , FSH and LH increase. During these peri-menopausal years, the ovulatory and menstrual cycles are irregular and the ovarian hormones no longer respond to stimulation from the gonadotropins. Menopause can be diagnosed after a year of amenorrhea (no menstruation) and the diagnosis can be confirmed if serum levels of FSH exceed a certain level.[42]

2.2 Physiological consequences of menopause

2.2.1 Short term effects

Dennerstein et al[7] studied 172 Australian women from perimenopause to postmenopause (the climacteric period) to identify the symptoms they experienced during this period. They tracked the women annually and used questionnaires to assess which symptoms changed in prevalence and severity during midlife and which factors best explained the changes recorded in these symptoms. They identified hot flashes, night

sweats (vasomotor symptoms) and vaginal dryness (genitourinary symptoms) as symptoms related to the hormonal changes of menopause. Other symptoms often attributed to menopause, such as depression and nervous tension, showed no change in frequency before, during or after the climateric period.

Vasomotor symptoms are attributed to declines in estrogen levels and increases in LH associated with menopause. The exact mechanism is still not clear but it appears to originate in the hypothalamus—the thermoregulatory centre in the brain. If the temperature setting is for some reason lowered (below 37°C), the body will begin to sweat in an attempt to cool the core temperature to the new lower setting. Casper et al[42] argued that the combination of declining estrogen and increasing LH trigger the change in the hypothalamus, leading to hot flashes and night sweats. They also attributed the changes in the genitourinary system to declining estrogen levels. Blood flow to this heavily vascularized region is reduced because of estrogen deficiency and this leads to the atrophy of the vaginal epithelial tissue. This contributes to the dryness experienced as the membranes become thinner and hold less moisture.

2.2.2 Long term effects

Menopause can also have serious, long term effects on a woman's health as declining estrogen levels are thought to contribute to osteoporosis and heart disease. Osteoporosis is characterized by a loss of bone mass and the deterioration of the bone tissue. While bone loss begins before menopause, it occurs more rapidly once estrogen levels begin to decline. As bones become more fragile, the postmenopausal woman faces an increased risk of fractures, particularly in the hip, spine and wrist areas.[43]

According to the Osteoporosis Society of Canada, there are currently 1.4 million Canadians with osteoporosis. They state that 1 in 4 women over the age of 50 have osteoporosis. Of the 25,000 hip fractures that occur in Canada each year, 70% of these (n = 17,500) occur among individuals with osteoporosis. Hip fractures are very serious injuries, leading to death in up to 20% of the cases, and disability and loss of independence in 50% of the survivors. They also estimate that the health care costs associated with the disease (mainly long term hospital and chronic care costs) amount to $1.3 billion annually.

Declining estrogen levels associated with menopause are also thought to contribute to an increased risk of cardiovascular disease.[44] For instance, a decrease in estrogen leads to changes in lipid metabolism which in turn contributes to adverse changes in serum cholesterol levels. A healthy serum profile, protective against the development of atherosclerosis, would include higher levels of high density lipoprotein (HDL) cholesterol and lower levels of low density lipoprotein (LDL) cholesterol and triglycerides. Declining estrogen levels lead to decreasing levels of HDL cholesterol and increasing levels of LDL and thus may contribute to an increased risk of heart disease and stroke.

Heart disease and stroke are the leading causes of death among Canadian women. According to the Canadian Heart and Stroke Foundation, cardiovascular disease accounted for 39% of deaths (approximately 42,000) among women in 1997. They also estimated the health care costs associated with heart disease at $18 billion annually.

Given the significant impact on health and high health care costs associated with the long term consequences of menopause, it is not surprising that therapies were actively

sought to help manage these consequences. Because decreased estrogen appears to contribute to menopausal symptoms, supplemental estrogen in the form of ERT is a standard treatment for postmenopausal women.

2.3 Estrogen replacement therapy

2.3.1 Benefits

To counteract the short term consequences of declining estrogen levels, physicians have been prescribing ERT to their postmenopausal patients since the late 1930's. Claims that ERT could control the menopausal symptoms led to an increase in prescriptions in the 1960's and early 1970's. According to an American pharmaceutical research database, the National Prescription Audit (NPA), 1975 was the peak dispensing year for ERT, with over 28 million prescriptions.[4] There are over 20 different ERT formulations. The most commonly prescribed form of estrogen is Premarin, a conjugated equine estrogen that is derived from the urine of pregnant horses.

There is strong evidence that ERT is the most effective treatment for the short term symptoms of menopause. One of the aims of the PEPI Trial was to assess the effects of estrogen on relief from short term menopausal symptoms. In a multi-center, double blind, placebo controlled trial involving 875 postmenopausal women, Greendale et al[11] reported that women on ERT experienced significant reductions in vasomotor symptoms compared to the placebo group. In an earlier study[45], daily ERT use was also reported to reduce the genitourinary symptoms associated with menopause.

There is also strong evidence that ERT assists in the maintenance of bone density, and thus can be used to prevent osteoporosis. In a meta-analysis involving 22 RCTs,

Torgerson et al[46] found a 27% reduction in nonvertebral fractures among postmenopausal women on HRT compared to postmenopausal women on placebo. There were also several small RCTs that found an increase in bone mass among women on ERT and a decrease in bone mass among the placebo group.[12] Nonrandomized studies have also revealed that postmenopausal women on ERT have a 25-30% reduced risk of hip fracture and a 50% reduced risk of vertebral fracture.[47]

The evidence supporting HRT use for reducing the risk of cardiovascular disease is not as clear. The results from over 30 nonrandomized studies suggest that women on ERT have a lower risk of heart disease compared to women not taking ERT.[8] Researchers involved in the PEPI Trial reported that levels of HDL and LDL cholesterol were more favorable among women on ERT compared to women on placebo.[14] However, results from more recent RCTs examining short term HRT use (four years or less) do not support these findings.[18] For instance, in the Estrogen Replacement and Atherosclerosis (ERA) Study, researchers reported no change in the coronary artery atherosclerosis among women on EPRT, ERT or placebo.[48]. In the secondary prevention Heart and Estrogen/progestin Replacement Study (HERS), researchers found no difference in the rates of nonfatal myocardial infarctions and coronary deaths among women on EPRT and those on placebo.[19]

There are other purported health benefits realized by postmenopausal women on ERT. In addition to osteoporosis and heart disease, postmenopausal women are also believed to have a reduced risk of developing Alzheimer's Disease. Several studies have reported that ERT may reduce the risk of dementia among postmenopausal women. For instance, Kimura[49] reported that women on estrogen demonstrated improved cognitive

function. Kampen and Sherwin[50] found that performance on memory tests was better among women on ERT. The results from other studies examining cognitive outcomes were not consistent with these findings and therefore it is difficult to draw any conclusions. Yaffe et al[21] performed a systematic review and meta-analysis on the relationship between dementia and ERT and concluded that adequate trials must be completed before any reasonable conclusions can be reached. In a similar type of study, Leblanc et al[51] reported a decreased risk of dementia among women on HRT, but argued that many of the studies included in their meta-analysis lacked methodological rigour.

There is also some evidence that women on ERT have a lower risk of colorectal cancer. In an early meta-analysis based on 14 studies, MacLennan et al[22] found that there was no association between ERT and colon or rectal cancer. However, Fernandez et al[52] concluded that more recent evidence supported a reduced risk of colorectal cancer among HRT users. Nanda et al[20] performed a meta-analysis of 25 nonrandomized studies and found a 33% reduction in colon cancer among recent HRT users. However, there was no difference in the risk of rectal cancer among HRT users and nonusers. The authors also concluded that the results of a RCT, part of the Women's Health Initiative due to come out in 2006, will provide further insight into this issue.

2.3.2 Risks

2.3.2.1 Breast cancer

In spite of the benefits associated with ERT use, there are also a number of health risks. One area of particular concern is the relationship between ERT use and breast cancer. Many postmenopausal women are reluctant to take estrogen because of a

perceived increased risk of breast cancer associated with its use.[53] There is already an increased risk of breast cancer among Canadian women once they reach age 50 [32] and the general fear is that the risk will increase further if they take ERT.

Breast cancer affects women throughout the world and is one of the leading causes of cancer deaths among women.[54] In Canada, breast cancer is currently the second most common cause of cancer death among women, after lung cancer.[32] There are many risks factors associated with breast cancer, including increasing age, family history, nulliparity and late age at first birth.[55] Many of the risk factors for breast cancer are related to estrogen exposure. Women who experience an early menarche (before age 12) or a late menopause (after age 55) and thus have a prolonged exposure to estrogens are at a higher risk compared to women with late menarche or early menopause.[56] Exposure to hormonally active agents, such as the organochlorides DDT (dichlorodiphenyltrichlorethane) and PCBs (polychorinated biphenyls) is also believed to increase the risk of breast cancer.[57]

There have been numerous studies that have investigated the effects of estrogens on laboratory animals and *in vitro* systems. Estrogen has been found to induce breast cell proliferation and lead to a high incidence of breast cancer among rodents.[58] In studies of human breast tissue, the proliferative effects of estrogen is also well documented.[59;60] However, it is not clear whether this increases the risk of breast cancer among women on ERT.

Breast density is a strong predictor of breast cancer[61]. Boyd et al[62] found that women with high mammographic density had a four to six times higher risk of breast cancer compared to women with lower mammographic density.[62] A RCT was recently

completed that examined the effect of ERT on breast density. Freedman et al[63] found that women on ERT had significantly greater breast density than women receiving placebo. However, the study did not assess if the women on ERT had a higher risk of breast cancer.

There have been over 50 nonrandomized studies that have examined the association between ERT and breast cancer.[31] The results of these studies are somewhat contradictory, likely as a result of the variation between studies in terms of the type, dosage and duration of estrogen used. In addition, there have been six meta-analyses completed in an attempt to determine if women on ERT have a higher risk of breast cancer.[26-31]

The results from the meta-analyses were consistent when the risk of breast cancer was compared between two groups of postmenopausal women: ever users and never users. Ever users included those women that had been on ERT. The time frame for use was usually not specified but women who reported sporadic use were excluded from the analysis[29;30]. Never users consisted of postmenopausal women that had never used HRT. All of these meta-analyses reported no increased risk of breast cancer among ever users compared to never users. Three of the meta-analyses[27-29] found that the women who were exposed to higher daily doses of ERT (1.25 mg/day) had an increased risk of breast cancer. Duration of ERT use was also found to affect the breast cancer risk. Three of these meta-analyses[28;30;31] reported an increased risk of breast cancer among women with increasing duration (more than five years) of ERT use.

The scientific quality of these meta-analytic reviews was assessed using the Oxman and Guyatt scale.[64] This scale consists of 10 criteria designed to evaluate the

quality of the search methods and selection criteria, the validity of the included studies, the statistical methods used for combining studies and the conclusions. Only one of the breast cancer/ERT reviews relied on more than one electronic database to search for articles[28] while the others used Medline only. None searched EMBASE. One[26] was completed before the advent of such databases. One[31] relied on individual patient data rather than the summary risk estimates from published studies. None of the reviews used unpublished literature while two[27;28] specifically mentioned ignoring this body of work. There was considerable variation between these six reviews in terms of the inclusion criteria that were used to select studies, and all but one[26] described the criteria that they used. Only three of these reviews mentioned assessing the quality of studies that were included[27-29], but none reported using an established quality assessment tool. The types of statistical analyses varied somewhat, but most used the fixed effects model and one[30] used the random effects and fixed effects model. Overall, all but one[26]of the reviews were considered to be well done, having only minor or minimal flaws. Therefore, the conclusion from these meta-analytic reviews, that ERT use does not increase the risk of breast cancer among postmenopausal women, is considered to be reasonable.

2.3.2.2 Endometrial cancer

As the number of women on ERT increased in the early 1970's, so did the incidence of endometrial adenocarcinoma among women with an intact uterus.[65] In Canada, the ASIR for endometrial cancer among postmenopausal women increased 50% (from 60 per 100,000 in the early 1970's to over 90 per 100,000 in the late 1970's).[33]

There have been over 35 nonrandomized studies that examined the relationship between ERT use and endometrial cancer. In a meta-analysis that summarized the results of these studies, Grady et al[34] reported that estrogen duration and dosage influenced the risk of this relatively rare cancer. Long term users of ERT (greater than 10 years) had an increased the risk of endometrial cancer 10 times that of women not taking ERT. Women on ERT for one to five years had a three fold increased risk. Their risk also increased when exposed to higher doses of estrogen[66].

Laboratory studies revealed that estrogen stimulates the proliferation of the endometrium. Continued stimulation, such as that experienced with long term ERT use, leads to hyperplasia, which can then develop into cancer.[67]

However, the addition of progestins (synthetic progesterone) to the estrogen formulation reduced the risk of endometrial cancer. Whitehead et al[68] examined the effects of estrogen and progestin on the endometrium. They found that progestin inhibited the proliferative effects of estrogen and thus prevented endometrial hyperplasia. Several epidemiological studies have reported no increased risk of endometrial cancer associated with EPRT.[69;70] While women that have had a hysterectomy continue to receive ERT, most doctors now prescribe EPRT to non-hysterectomized postmenopausal women so they will still realize the benefits of hormone replacement therapy without the added risk of endometrial cancer. The ASIR of endometrial cancer among postmenopausal women in Canada has declined steadily since their peak in the late 1970's. The most recently reported ASIR of endometrial cancer is approximately 64 per 100,000, similar to the rates reported in the early 1970's.[33]

2.4 Estrogen progestin replacement therapy

The use of EPRT gradually increased from the early 1980's to the 1990's. According to the Nurse's Health Study, there were very few women taking EPRT prior to 1986.[71] In 1986, they reported that 18% of the registered nurses in the study were on EPRT. EPRT use had increased to 30% among this group by 1990. It is clear from laboratory and epidemiological studies that the addition of progestins reduces the risk of endometrial cancer[34]. However, it is still not clear what effect EPRT use has on breast cancer. Early studies reported that progestins are protective against breast cancer.[36;37] More recent studies suggest that progestins increase the risk.[39;40]

In the endometrium, progestins oppose the proliferative effects of estrogens and thus reduce the risk of endometrial cancer. However, the relationship between progestins and breast tissue is more complex. Some studies report that progestin stimulates cell proliferation, while others report that progestin's action is inhibitory.[72] This may be due in part to the fact that there are two types of progestins that are used in EPRT: progesterone derived compounds and testosterone derived compounds. Depending on their structure and location, these compounds can behave in a variety of ways that range from estrogenic to androgenic.

The results from the PEPI Trial demonstrate the increased proliferative effects of EPRT. Greendale et al[73] monitored changes in breast density change among women on ERT, EPRT and a placebo group (no HRT). Breast density is generally believed to be a strong predictor of breast cancer.[61] After 12 months, the researchers reported that 19.4% of the women on EPRT and only 3.5% of the women on ERT had an increase in breast density. There was no increase in breast density among women in the placebo group.

Several of the meta-analyses examining the relationship between ERT and breast cancer also examined the association between EPRT and breast cancer. According to Sillero-Arenas et al[29], there was no increased risk of breast cancer associated with EPRT use, based on the results of three nonrandomized studies (OR = 0.99; 95% CI 0.72-1.36). Using four nonrandomized studies studies, Colditz et al[30] reported similar findings (OR = 1.13; 95% CI 0.78-1.64). Beral et al[31] found no differences in the risk of breast cancer among women on ERT or EPRT. Since these meta-analyses have been completed, there have been several additional studies that have examined the relationship between EPRT use and breast cancer in postmenopausal women. Because recent research suggests an increased risk of breast cancer associated with EPRT use[38-40;74], an investigation examining all of the breast cancer/EPRT research is needed in order to better understand the breast cancer risk associated with EPRT use. In a recent editorial, Dixon argued that the effects of EPRT on breast cancer risk needs to be considered, especially in light of the current evidence.[75] The aim of this study is to conduct a systematic review and meta-analysis of the literature incorporating all of the studies completed to date in order to examine the association between EPRT and breast cancer risk among postmenopausal women.

2.5 Systematic reviews and meta-analyses of nonrandomized studies

2.5.1 Rationale and definition

Klassen et al[76] report that, given the tremendous volume of health care literature published each year, systematic reviews of the literature assist health care professionals in keeping up to date with the latest results. A systematic review is defined as a "*review*

in which there is a comprehensive search for relevant studies on a specific topic and those identified are then appraised and synthesized, according to a predetermined and explicit methodology".[77] Egger and Smith[78] use the term systematic review to refer to "*any type of review that has been prepared using strategies to avoid bias and that includes a material and methods section; it may or may not include formal meta-analysis*". The systematic review to be performed for this thesis will follow these definitions.

Greenland[79] distinguished between the two goals of a systematic review: synthetic and analytic. A review that focuses on the "synthetic" goal is the classic meta-analysis, defined as a "*quantitative approach for systematically assessing the results of previous research in order to arrive at conclusions about the body of research.*"[80] This ultimately provides an estimate of an overall summary or average effect across studies. A review that focuses on an analytic goal leads to the identification and estimation of differences among study specific effects and can produce a better understanding of the relationship between the outcome and the exposure of interest.

The major goal of this systematic review will be analytic in nature, with summary estimates determined where appropriate. This approach will permit a methodical way of analysing studies and identifying patterns among study results.[81] It will also allow for the identification of possible reasons for variability and inconsistency between nonrandomized studies.

2.5.2 Cochrane collaboration and Cochrane systematic reviews

The Cochrane Collaboration was founded in 1993. Its current mandate is to "*prepare, maintain and promote the accessibility of systematic reviews of the effects of health care interventions*". Its development stemmed from the work of Archie Cochrane, a physician and epidemiologist interested in health care effectiveness and efficiency. He stressed the importance of using evidence from RCTs to make decisions on health care needs.

There are currently 50 different Collaborative Review groups that are responsible for conducting systematic reviews in specific research areas (eg. breast cancer research group, musculoskeletal group). All systematic reviews prepared by these groups follow the recommendations provided in the Cochrane Collaboration Handbook[82]. This handbook is updated quarterly and provides specific details on the steps needed to prepare and maintain a systematic review. The most recent version of the handbook is readily accessible on their website (www.medlib.com/cochrane).

Jadad et al[83] compared the methodology and reports of systematic reviews and meta-analyses that were prepared using Cochrane methodology with non-Cochrane reviews. They reviewed 36 Cochrane reviews with a random sample of 39 non-Cochrane reviews published since 1995. They concluded that Cochrane reviews had greater methodological rigor and were less prone to bias.

The methodology outlined in the most recent Cochrane Reviewer's Handbook[82] was applied when preparing this systematic review. While the Cochrane Collaboration focuses on evidence obtained from RCTs, they also provide additional information pertaining to nonrandomized studies that were incorporated into this current review.

One of the problems associated with systematic reviews and meta-analyses identified in Jadad et al[83] concerns the quality of their reporting. In at attempt of improve this, a Quality of Reporting of Meta-analyses (QUORUM) group established a checklist of reporting standards.[84] This 17 item checklist focuses primarily on reporting evidence from RCTs.

To address the specific concerns of nonrandomized or observational studies, another group, the Meta-analysis of Observational Studies in Epidemiology (MOOSE) also developed a checklist for reporting the results of meta-analyses.[85] This 35 item checklist was developed through the deliberations of 27 participants (the MOOSE Group) and provides specific details about what should be included in the reporting of a meta-analysis involving nonrandomized or observational studies. Therefore, in addition to following the guidelines provided in the Cochrane Manual, the QUORUM and MOOSE checklists were also used in the preparation of this report to assess completeness of the reporting of the studies included.

2.5.3 Criticisms and concerns of the meta-analysis

Despite its widespread use, meta-analysis continues to be a controversial technique. For instance, Jadad et al[86] reported that meta-analyses addressing the same therapeutic question often reach opposite conclusions. Such discordance will present difficulties for decision-makers who rely on these reviews to help them make choices among alternative health care interventions. Egger and Smith[87] argued that these discrepancies are due to various types of selection bias inherent in the studies chosen for

inclusion in the meta-analysis. Detsky et al[88] proposed that variation in the quality of the trials that are included in the meta-analysis can also lead to bias.

Even more controversial is the application of meta-analytic techniques to nonrandomized studies. The main problem with nonrandomized studies involves the composition of the study groups. In an RCT, the clinicians can allocate participants into treatment and control groups randomly and with concealment. When designing nonrandomized studies, the participants are assigned to their groups based on their prior "experiences" and "choices". In cohort studies, they are either exposed or unexposed to the intervention. In case control studies, they already have the outcome (case) or not (control).

However, there are situations where randomized studies are not always appropriate and the only option remaining is nonrandomized studies. For instance, it is unethical to conduct an RCT when the exposure of interest is known to be hazardous. There have been many RCTs which have looked at the relationship between hormone replacement therapy and osteoporosis but none to date that have focussed primarily on HRT and breast cancer. All studies that have cancer as a primary outcome are nonrandomized. Rather than ignore the findings of these studies, it is important to develop more rigorous methods to assess the evidence. One way to accomplish this is by assessing the quality of studies that are incorporated into the meta-analysis.

2.5.4 *Assessing the quality of studies in a systematic review/meta-analysis*

2.5.4.1 Importance

There have been two studies that have evaluated the role that quality assessment plays in meta-analytic results.[10:89] In both studies, the authors found that the inclusion of poor quality studies in the analysis altered the results considerably. By assessing the quality of a study, the researcher is provided with an indication that the results are a more valid estimate of the truth and thus should be incorporated into a meta-analysis.[90]

Petitti[80]argued that study quality can serve as an indicator of study validity. She stated that it is important to assess the quality of studies used in a meta-analysis. By taking into account study quality, the validity of the meta-analysis can be enhanced.

2.5.4.2 Evaluation of quality assessment tools

Numerous scales and checklists have been developed to assist with the quality assessment of studies, but most of these scales have been developed specifically to assess the quality of RCT. The current gold standard for RCT evaluation is the Jadad scale[90]. There are several scales that have been developed to assess the quality of nonrandomized studies, but there is no current gold standard.

A comprehensive review of quality assessment tools for nonrandomized studies was completed by Dinnes et al.[91] They performed a thorough review of the literature that included a search of 10 different electronic databases (including Medline, Embase, PsycLit); a review of abstracts and reference lists and advice from experts. Their search yielded over 7000 articles, and from these, they identified 210 potential quality assessment tools. They examined each of these tools and selected those with more than

five internal validity domains and more than three core items. They identified six of these tools as having met their inclusion/exclusion criteria and being suitable for use in systematic reviews.

Most of the studies that were included in the systematic reviews of ERT and breast cancer were either case control or cohort. Based on preliminary findings, the same types of study have been used to study the relationship between EPRT and breast cancer. Therefore, it is essential that the assessment tool can be used to evaluate both types of studies. The Newcastle Ottawa Scale (NOS)[92] was the only tool identified in the Dinnes review that can be used to assess case control and cohort studies. The others, two of which remain unpublished, were designed to assess quality in cohort studies or cohort and RCTs [93-95]

2.5.4.3 The Newcastle Ottawa Scale

2.5.4.3.1 Description

The Newcastle Ottawa Scale (NOS) consists of two different scales; one designed specifically for cohort studies and the other for case control studies (Appendix A). In addition to assessing study quality (in terms of representativeness of the participants, exposure and outcome assessment), this scale also permits an assessment of confounding.

Each NOS is composed of eight items that enable the reviewer to assess how well the study was performed in terms of controlling potential bias. It evaluates a study based on three criteria that reflect the major categories of bias that are often associated with nonrandomized studies: Selection of the study groups; Comparability of the two groups of individuals (exposed and nonexposed in the cohort studies; cases and controls in the

case control studies); and Ascertainment of outcome (cohort studies) or exposure (case control studies). These criteria enable the evaluator to: assess the selection of individuals in the study (are they representative of the community?); evaluate the method of exposure (pharmacy records; self report? questionnaire?) and outcome ascertainment (pathology report?); and examine the comparability of the cases/controls or exposed/unexposed individuals (same age? same parity?) which are involved in the study.

Each criteria has two to four questions in order to assess how well the study deals with bias and confounding. The scoring consists of a star system; stars are awarded if the study meets a particular criteria featured in the item; the more stars awarded then the higher the study quality. For instance, if the exposure was assessed through medical records, a star is awarded but if it was obtained from a written self report, no star is given. Another important component, particularly in cohort studies, is the loss to follow-up. If the study is able to keep track of 80% or more of its original participants, the NOS will award a star. If more than 20% of the participants are lost to follow-up, no star is awarded.

Because the participants in nonrandomized studies are not randomly allocated to the study groups, it is essential that the major risk factors (such as age) are controlled. NOS awards two points if two of the major risk factors are comparable in cases and controls groups. A study can gain a maximum of four stars for Selection characteristics; two stars for Comparability and three stars for Ascertainment of exposure/outcome, for a maximum score of nine.

The validity and reliability has been formally assessed in only one[94] of the six tools identified by Dinnes et al[91]. This tool, developed by Black and Downs, is a

comprehensive scale that provides an overall score for quality for each study. However, only 18 of its 27 items were appropriate for assessing the quality of a cohort study and even fewer items could be applied to case control studies.

A formal evaluation of the NOS has been initiated (Appendix B). Preliminary findings suggest that the NOS has good inter-rater and intra-rater reliability (Intra-class correlation ≥ 0.84). When compared with the Black and Down's scale, the criterion validity of the NOS was also very good (Intra-class correlation ≥ 0.65). To date, the NOS has been used to evaluate study quality in several meta-analyses, including the association of breast cancer with estrogen replacement therapy[92]; the effect of arterial revascularization on survival[96]and the association of rheumatologic disorders with silicone breast implants[97].

2.5.4.3.2 Assessment of bias

The Cochrane Collaboration recommends that the critical appraisal of studies should focus on how well particular studies control for bias.[82] They suggest four types of bias to consider when performing quality assessment of nonrandomized studies: selection, performance, attrition and detection bias, and they provide criteria that can be used to assess each of these biases (Table 2.1).

Selection bias. One of the main differences between RCTs and nonrandomized studies involves the selection of participants. In RCTs, this problem can be resolved through allocation concealment. This is not possible in nonrandomized studies.

Table 2.1. Criteria used to critically appraise the quality of nonrandomized studies based on the four main sources of bias identified in Clarke and Oxman[82]

Source of Bias	Cohort studies	Case control studies
Selection bias	Control for confounders	Matching
Performance bias	Measurement of exposure	Measurement of exposure
Attrition bias	Completeness of follow-up	Non-response rate
Detection bias	Blinding	Case definition

Selection bias is one of the most important biases as it may distort comparisons between cases and controls in case control studies and exposed and unexposed individuals in cohort studies. When assessing how well nonrandomized studies deal with selection bias, reviewers must identify what confounders they deem to be important and assess the extent to which these confounders were measured and controlled for in the studies.

Performance bias. In order to assess the performance bias in nonrandomized studies, the methods used to measure exposure must be similar in the groups being compared. For this systematic review, it is essential that the measures of exposure to EPRT therapy are similar among the cases and controls in case control studies and among the exposed group in the cohort studies.

Attrition bias. This type of bias occurs when there are systematic differences between groups (cases and controls; exposed and unexposed) in terms of dropouts or losses of participants. It is essential to assess how a particular study deals with these losses. If these are not properly accounted for, it has the potential to bias the results.

Detection bias. Systematic differences in outcome assessment can result in detection bias. To avoid this bias, a clear case definition must exist in case control studies and a blind assessment of outcome must appear in cohort studies.

The eight items in the NOS (Appendix A) clearly address these four types of biases. Table 2.2 illustrates which items in the NOS specifically address each type of bias. Therefore, evaluating a study using the NOS would permit the reviewer to determine how well each type of bias was controlled. A high NOS score would indicate that these biases were addressed and thus the study would be considered to be of higher quality than one where the bias was not controlled.

Table 2.2. Types of bias examined by specific questions in the Newcastle Ottawa scale

Type of Bias	NOS for case control studies	NOS for cohort studies
Selection bias	Items 3,4 and 5	Item 5
Performance bias	Items 6 and 7	Items 1, 2 and 3
Attrition bias	Item 8	Items 7 and 8
Detection bias	Items 1 and 2	Items 4 and 6

3. METHODS

In this chapter, the search strategy employed to identify the pertinent literature will be described and the electronic databases used in the search will be profiled. An explanation of the methods used to screen and select articles, including the eligibility criteria, will be presented and details about the quality assessment process using the Newcastle Ottawa Scale will be reviewed. A description of the fixed effects model and statistical testing that was used to determine the summary estimates and assess their significance will be provided. Other analytic strategies, including subgroup, sensitivity and cumulative meta-analyses will also be presented.

3.1 Search strategy

Felson[98] states that "the validity of a meta-analysis depends upon the complete sampling of all the studies performed on a particular topic". An extensive search of the literature was carried out to identify all studies that have examined the relationship between EPRT and breast cancer among postmenopausal women. The search strategy was developed with the assistance of information science specialists and librarians at the University of Ottawa. Several papers on developing strategies were also reviewed to gain additional information to further refine the search strategy.[99;100]

For this systematic review, a variety of sources were used including several electronic databases, hand-searching, review of abstracts from pertinent scientific meetings and review of literature cited in retrieved articles. Ten different databases were selected and search strategies were developed to examine each database. A description of each database and rationale for its use is provided below.

One of the largest databases used was Medline. It is produced by the U.S. National Library of Medicine and covers the international biomedical literature. It indexes 4000 journals that are published in the United States and 70 other countries. It also provides records that are indexed by HealthStar. While most of the citations are from English language journals, Medline does include references in other languages (e.g. French, German, Italian). This database contains records that are indexed by *Index Medicus.* Pre-Medline was also searched. This database provides bibliographic details about recently published studies that have not as yet been indexed by Medline.

Cancerlit is produced by the U.S. National Cancer Institute and covers all aspects of cancer. It is international in scope but mainly focuses on English language publications. While some of the citations in Cancerlit overlap with Medline, it also covers information on cancer-causing agents and the mechanisms of mutagenesis that would be pertinent to this search. In addition to reviewing 200 core journals, it also indexes scientific meetings and government reports, thereby providing some access to the grey literature.

Current Contents reviews 7,500 international periodicals and provides citations for journal articles, reviews, conferences and editorials. Further international coverage was also provided by Biological Abstracts. It indexes medical research findings and clinical studies. Dissertation Abstracts was also searched because it provides a potential source of unpublished material by indexing Masters and Doctoral Theses.

Embase is considered to be the "European Medline". It provides an electronic version of *Excerpta medica.* This database indexes more than 1,000 journals that are not found in Medline[100]. Smith[101] reports that the overlap in journals that are indexed by

Medline and Embase is approximately 34%, but it varies depending on the subject area. This indicates that searching Embase yields articles not found in Medline/Cancerlit.

The basic search strategy employed for these seven electronic databases consisted of three main concepts that were searched separately: (i.) breast cancer studies (sample MeSH term: breast neoplasms); (ii.) studies on HRT (sample MeSH term: estrogens, conjugated); and (iii.) types of studies (sample MeSH term: case control studies which encompasses nested and matched case control studies). These three sub-searches were subsequently combined using the Boolean operator "OR". Use of the breast cancer and HRT MeSH headings and "OR" would identify any studies indexed on at least one of these headings. To ensure that all of the articles under a particular MeSH heading were retrieved, the subject heading was "exploded" by inserting "exp" before the heading. Since there were no MeSH headings to encompass "EPRT", specific text words or phrases were inserted into the search strategy. For instance, the text words "(estrogen progestin adj2 therap$).tw." would lead to a search of these words as free text in the titles and abstracts. The "adj2" implies that there can be up to two words between progestin and therapy (for instance *replacement* or *hormone replacement*). The "$" is a truncation that is used to permit alternative versions of the word, in this case, therapy or therapies.

The specific search strategy, developed for Medline/Pre-Medline (1966 to 10 October 2001) appears in Appendix C-1. A modified version (removing "Breast Cancer" as a text word) of this strategy was used to search Cancerlit (1975 to October 2001). In order to search Current Contents All editions (1993 Week 26 to 2001 Week 17), Dissertation Abstracts (1861 to October 2000) and Biological Abstracts (1990 to March 2001), MeSH headings were further modified (see Appendix C-2). For instance, when

searching for types of studies, the text words "odds ratio" was used instead of the MeSH heading "case control studies". No filters or limits such as year of publication or language were employed in any of these searches. These databases were accessed via the POLARIS network at the University of Ottawa, using the OVID vendor, Silver Platter version.

The search strategy was further modified in order to search Embase (Appendix C-3). The three main search areas: HRT; breast cancer and types of studies, were retained but some filters were added (human studies only, exclude studies with tamoxifen). The search was restricted to those articles that had been indexed from 1988 to present. From a preliminary search of the breast cancer/EPRT literature using Medline and Cancerlit, it was evident that most of the articles on EPRT use had been published after 1990. Embase provides different ranges of years that could be searched and the database preceding the 1988 to present included studies from 1980 to 1987. Three of the databases used in this study already indexed articles during this time frame and it was believed that these would be sufficient to capture any early articles about EPRT use. Colditz et al[71] reported that in 1986, only 18% of the postmenopausal women on HRT were using EPRT. Therefore, there was not widespread use of EPRT at this time. Since the costs associated with an Embase search can be quite prohibitive search ($2 per citation plus run time costs), the search was limited to those studies published after 1987.

Two additional electronic databases: PubMed and Biomedcentral were also explored. PubMed (http://www.pubmed.com) is produced by the National Library of Medicine. It is a search tool that enables the researcher to access the biomedical literature and is useful for gaining access to citations that precede the date for Medline

indexing. Another useful feature of the PubMed is that it provides a list of related articles for each of the citations identified during the search. Therefore, one can gain access to additional relevant articles by accessing the "Related Articles" feature in this database. Biomedcentral http://www.biomedcentral.com is the PubMed equivalent in Europe. It is also an independent publishing house that provides immediate access to 50 on line journals, including Breast Cancer Research. These databases provide current information, but the information can only be accessed using a very simple search strategy (eg. "HRT and Breast Cancer").

There are several different search engines that can access information on the World Wide Web. One of the best and most widely used of these is Google (http://www.google.com). It claims to have the most sophisticated search technology and thus produces more relevant results than any other directory search. It was used primarily as a source of general information and for access to current media coverage that could alert one to new reports in the literature. The same search strategy employed for PubMed and Biomedcentral was used when searching Google.

3.2 Screening and selection of articles

Once the references were revealed from all of the electronic searches, they were imported into a database within Reference Manager 9.0. This citation management software package assigns each imported article a unique identification number along with the bibliographic details of the article (author, title, year, journal, issue, pages, abstract, keywords). The title of each article was scanned and duplicates were removed. When

removing duplicates, the database source of each citation was recorded to gain some perspective of the degree of overlap between databases.

The titles and abstracts of each unique article were then examined to assess their suitability for inclusion in the systematic review. Two individuals (Joan Peterson, Loeb Research Institute; author) independently screened the title, abstract and key words for each citation by broadly applying the inclusion/exclusion criteria. A citation was retained within the Reference Manager database if the bibliographic details suggested that it may contain information on the relationship between exposure to HRT and breast cancer among postmenopausal women. All selected articles were retained within the Reference Manager database and deemed potentially relevant.

Four criteria were used to evaluate an article's suitability: the population under study, exposure, outcome and design (see Table 3.1). The population and exposure criteria pertained to postmenopausal women who have either been exposed to EPRT or have never been exposed to any postmenopausal HRT. The outcome of interest was breast cancer and the study designs of interest were RCTs, cohort and case control studies.

For RCTs, the treatment group consisted of postmenopausal women receiving EPRT and the placebo group was composed of postmenopausal women not receiving EPRT and having never been exposed to any HRT. For cohort studies, the exposed group consisted of postmenopausal women receiving EPRT during the follow up period and the unexposed group was the postmenopausal women who were hormone free for the entire follow up. The endpoint for these studies was incident breast cancer. For case control studies, cases consisted of all postmenopausal women that were diagnosed with

breast cancer and controls were all postmenopausal women with no history of breast cancer. The endpoint was the history of postmenopausal EPRT.

Table 3.1. Eligibility criteria used in the selection of studies to examine the association between breast cancer risk and EPRT use among postmenopausal women.

Characteristic of study	Inclusion criteria	Exclusion criteria
Population	Postmenopausal women with no prior history of breast cancer	Premenopausal women; postmenopausal women with a history of breast cancer
Exposure	Exposed: current or past exposure to EPRT Unexposed: no current or past exposure to either EPRT, ERT or HRT	Women with current or past use of ERT or HRT (with no distinction between ERT and EPRT); Progestin-only replacement therapy; Estrogen/progestin contraceptive pills; Tamoxifen; Raloxifene
Outcome	Invasive breast cancer	Intermediate outcomes such as in situ cancer (lobular or ductal), changes in breast density as detected by mammography Final outcomes such as mortality from breast cancer
Design	RCTs, cohort and case control studies	Case reports, case series, ecological studies, cross sectional surveys

If the article satisfied these inclusion criteria, it was marked in the Reference Management file and later saved into a separate file for further assessment—the final stage of the screening process. If the article failed to meet the inclusion criteria, the reason for its rejection was recorded in a rejection log that was maintained throughout the entire screening process. For each article excluded, the reason for rejection was recorded along with the article's identification number (generated by the Reference Manager software). The total number of articles rejected and retrieved was recorded.

After this initial screening, the selected articles were reassessed by a secondary screening process (completed by the author) to determine suitability for inclusion in the systematic review and meta-analysis. This secondary screening process permitted the identification of multiple publications reporting on the same subjects. In this instance, all the publications were retrieved, but only the most recent analysis was included in the review. Attempts were made to contact the study authors for missing information (eg. standard errors or confidence intervals). If three attempts failed to provide the required information then the study was included for assessment in the systematic review but excluded from the quantitative analysis. A flow chart was developed to illustrate the steps involved in the entire screening process and the number of articles that were excluded at each step (Figure 4.1).

To ensure that recent citations not yet electronically indexed were included in this review, pertinent journals were handsearched (Table 3.2). This involved manually examining the current issues (all issues published in 2001 unless otherwise noted) of a particular journal and reading the title of each article to determine if it was relevant to the review. If the title appeared relevant, the abstract was read. The article was retrieved if it met the inclusion/exclusion criteria. The abstracts from the American Society of Epidemiologic Research were also searched from 1995 to 2001 to locate studies that were presented at the annual meetings of the American Society for Epidemiologists.

Table 3.2 List of handsearched journals. (All 2001 issues published up to October 2001 were handsearched, unless otherwise noted)

American Journal of Epidemiology; Annals of Epidemiology; Breast Cancer Research (1999 to present); British Journal of Cancer; Cancer: Detection and Prevention; Cancer Epidemiology, Biomarkers and Prevention (1999 to present); Cancer, Causes and Control; Climacteric (2000 to present); Epidemiology; European Journal of Cancer Prevention; European Journal of Epidemiology; International Journal of Cancer (1999 to present); International Journal of Epidemiology; Journal of Clinical Epidemiology;	Journal of Epidemiology and Community Health; Journal of the American Medical Association; Journal of the National Cancer Institute (1999 to present); Journal of Women's Health and Gender-based Medicine; Lancet; Lancet Oncology; Maturitas (1999 to present); Menopause (1997 to present); New England Journal of Medicine; Obstetrics and Gynecology; Women and Health.

The reference lists of all retrieved articles were also examined for pertinent studies and reviews. Once all of the screening was completed, a list of the studies to be included in the systematic review and meta-analysis was prepared. At this point, two individuals well versed in the field of HRT and breast cancer (Dr. A. Cranney, Ottawa Health Research Institute; Dr. E. Grunfeld, Ottawa Regional Cancer Centre and Ottawa Health Research Institute) were consulted. They were asked to review the list of studies to assess its completeness and see if they were aware of any additional pertinent published or unpublished studies that were omitted.

3.3 Assessment of the methodological quality of studies

The methodological quality of each study selected for inclusion in the meta-analysis was evaluated using the Newcastle Ottawa Scale (NOS)[92]. The NOS consists of two different scales, one to assess the quality of case control studies and the other for

cohort studies. A detailed description of the NOS appears in section 2.5.4.3.1. The quality of each study included in the meta-analysis was evaluated using NOS scoring sheets designed specifically for this investigation (Appendix A). An evaluation of the validity and reliability of the NOS appears in Appendix B.

3.4 Data abstraction

All pertinent data from each of the studies revealed through electronic databases, handsearching or reference list review was abstracted using standardized Data Abstraction forms (Appendix D). These forms were designed to systematically extract comparable qualitative and quantitative information from each of the selected studies. The information obtained from each study included specific details about the study design, analysis and conclusions. Table 3.3 provides an overview of the information that was abstracted.

Table 3.3. Information abstracted from selected studies

Study details: • *Author(s)* • *Year of publication* • *Language of publication* • *Country of study* • *Years of study* • *Aim of study* • *Type of study* • *Study outcome* • *Criteria used to assess outcome* Description of EPRT • *Formulation* • *Dosage of estrogen and progestin* • *Classification of exposure* • *Number of days progestin taken*	Description of Cases and Controls/Cohort • *Number excluded and explanation* • *Group determination* • *Ascertainment of exposure* Subject characteristics Group differences Analysis: • *Number of exposed cases/controls* • *Number of unexposed cases/controls* • *Risk factors controlled in design* • *Risk factors controlled in analysis* • *Adjusted odds ratios ± confidence limits* • *Adjusted relative risks ± confidence limits* • *Person years of observation* Concerns/comments

3.5 Quantitative data synthesis

The fixed effects model was used to analyze the data extracted from each of the selected studies. There are several software packages (eg. RevMan 3.0; Meta-Analyst®) that have been developed to analyze the data for a meta-analysis. However, they are designed specifically for the analysis of randomized control trials and cannot be used for nonrandomized studies. Therefore, the summary measures and critical values for chi-square tests of significance were computed using Microsoft Excel (Version: Windows 1997) with calculations based on the fixed effect model formulas in Fleiss[102].

Summary measure: The adjusted odds ratio (OR) or relative risk (RR), and their respective 95% confidence intervals (95% CI), were entered on the spreadsheet and each of these three values was log transformed (see Appendix E for a sample spreadsheet). The log transformed values of the OR/RR represent the Y_i 's in the fixed effect model equations. The standard error (s.e), "s_i" was calculated from the confidence intervals using the formula:

$$\text{s.e.} = \frac{(\text{Upper CI} - \text{lower CI})}{2(Z_{\alpha/2})}$$

where α was set at 0.05 and thus $Z_{\alpha/2} = 1.96$. The weight of the estimate W_i, was calculated by taking the inverse of the standard error squared.

$$W_i = 1/s_i^2$$

The product of the log transformed measure Y_i and its weight, W_i. were then determined and the sum of these products divided by the sum of the weights (ΣW_i.) provided an estimate of the overall measure, θ:

$$\theta = \frac{\Sigma W_i Y_i.}{\Sigma W_i.}$$

χ^2 statistical tests: The critical values for the χ^2 were calculated using the same spreadsheet and the columns of values entered earlier. The three different χ^2 values were calculated using the following formulas from Fleiss[102].

$$\chi^2_{\text{total}} = \Sigma W_i Y_i^2$$

$$\chi^2_{\text{association}} = (\Sigma W_i Y_i)^2 \approx \chi^2_1$$

$$\chi^2_{\text{homogeneity}} = \Sigma W_i (Y_i - \theta)^2 \approx \chi^2_{k-1}$$

The critical value in the χ^2 test for association was calculated to determine if there was an association between exposure to EPRT and breast cancer. The critical value in the χ^2 test for homogeneity, Cochran's Q, was calculated to determine if there was significant heterogeneity between the studies entered into the meta-analysis. If this χ^2 value is significant, then heterogeneity was further explored. Prior to using a random effects model, which is often used when there is heterogeneity between studies, the heterogeneity was explored in more detail using subgroup and sensitivity analyses.

There are several factors that could contribute to the variation between studies. These include: study quality; differences in the duration of EPRT use (long vs short term use); recency of use (current users); the geographic location of the study (Europe or North America); and type of EPRT (a progesterone or testosterone derived progestin). Each of these factors was explored in the subgroup analysis.

Sensitivity analyses are designed to determine how sensitive the conclusions are to the various assumptions made in the analysis. According to the Cochrane Reviewer's Handbook[82], this method of analysis provides "an approach to testing how robust the

results of the review are relative to key decisions and assumptions that were made in the process of conducting the review". If there was uncertainty whether a particular study met the eligibility criteria (eg. studies lacking adjusted risk estimates) a sensitivity analysis was performed to determine what effect the inclusion or exclusion of this study had on the summary estimate and the statistical analysis.

Cumulative meta-analyses were also performed as a type of sensitivity analysis to determine if the results varied according to particular characteristics such as year of publication or study quality. Cumulative meta-analyses were first developed by Lau et al[103] to track trends in therapeutic advances and to see if and when research of a new treatment is sufficient to determine whether it is effective or ineffective. In this study, a cumulative meta-analysis was used to determine if the results varied over time and with studies of different quality. The chronological cumulative meta-analysis involved adding studies one at a time starting with the earliest study and ending with the most recently published. For the study quality cumulative meta-analysis, the highest quality studies were entered first, followed by studies with lower quality scores.

After all of the studies were placed in the order of interest, the OR or RR of the first study was entered into the analysis as the first point. The second entry was the combined summary estimate of the first and second study, calculated using the fixed effects model and the Microsoft Excel worksheets. This process continued in a sequential fashion with cumulative estimates calculated at each step until all studies were incorporated into the analysis. If the summary estimate remained constant with each additional study added, it would suggest that the results have been robust over time or across studies of differing quality. If the estimates became progressively smaller or

larger with changes in study quality, it would suggest that the results are not robust across studies of different quality.

The results of the subgroup and cumulative meta-analyses were graphed onto Forrest Plots using the high/low/close graphs provided in Harvard Graphics Version 3.0. In these plots, the individual odds ratios/relative risks and their 95% CI are log transformed. The 'high' value corresponded to the upper confidence limit; the 'low' value represented the lower confidence limit and the close was the individual odds ratio or relative risk. The summary measure is entered at the bottom of the plot with its respective 95% CI. A line representing no effect (OR or RR = 1) is also plotted. Summary estimates that are plotted on the right of this line indicated that the risk of breast cancer was higher with exposure to EPRT, while estimates to the left of this line indicated that there was a lower risk of breast cancer associated with exposure to EPRT. If either of the confidence intervals cross this line, it indicated that there was no significant association between breast cancer and exposure to EPRT.

In order to assess publication bias, funnel plots were constructed using the X-Y scatterplot option in Microsoft Excel 3.0. A funnel plot graphs the treatment effects from individual studies on the x axis against a measure of study precision on the y axis. Funnel plots enable one to assess whether small study effects have biased the findings of the meta-analysis.

When constructing funnel plots there are several options available for the vertical and horizontal axes. For estimates of study size, plotted on the vertical axis, one could use standard error, precision (1/SE), sample size or log sample size. For the horizontal axis, risk differences or ratio measures of treatment effect are plotted. Sterne and

Egger[104] evaluated these different choices by reviewing various meta-analyses and assessing how the choice of what is plotted on the x and y axes influences the shape of the funnel plot. They recommended using standard error as the measure of study size and log odds ratio as the measure for treatment effect. Clark and Oxman[82] also recommend the use of standard error as their estimate of study size.

Funnel plots were constructed by plotting standard error vs the log risk estimate (odds ratio or relative risk). In order to assess bias, the shape of the resulting plot was assessed to determine if it resembled the shape of an inverted funnel.

4. RESULTS

In this chapter, the results from the systematic review of the literature will be described. Reasons for the rejection and selection of articles for inclusion in the systematic review and meta-analysis will be provided. An assessment of the quality of these selected studies will also be presented. The results of a meta-analysis, performed to evaluate the relationship between breast cancer among ever users of EPRT versus never users of HRT will be reviewed. To further enhance our understanding of this relationship, the results of five different subgroup analyses (study quality, duration of EPRT use, current EPRT use, study location and EPRT formulation), will also be presented. To provide insight into the effects of including or excluding various studies from the analysis, the results of the sensitivity analysis and cumulative meta-analyses will be reviewed. The funnel plots will be examined in order to evaluate the presence of publication bias.

4.1 Summary of search results

4.1.1 Excluded studies

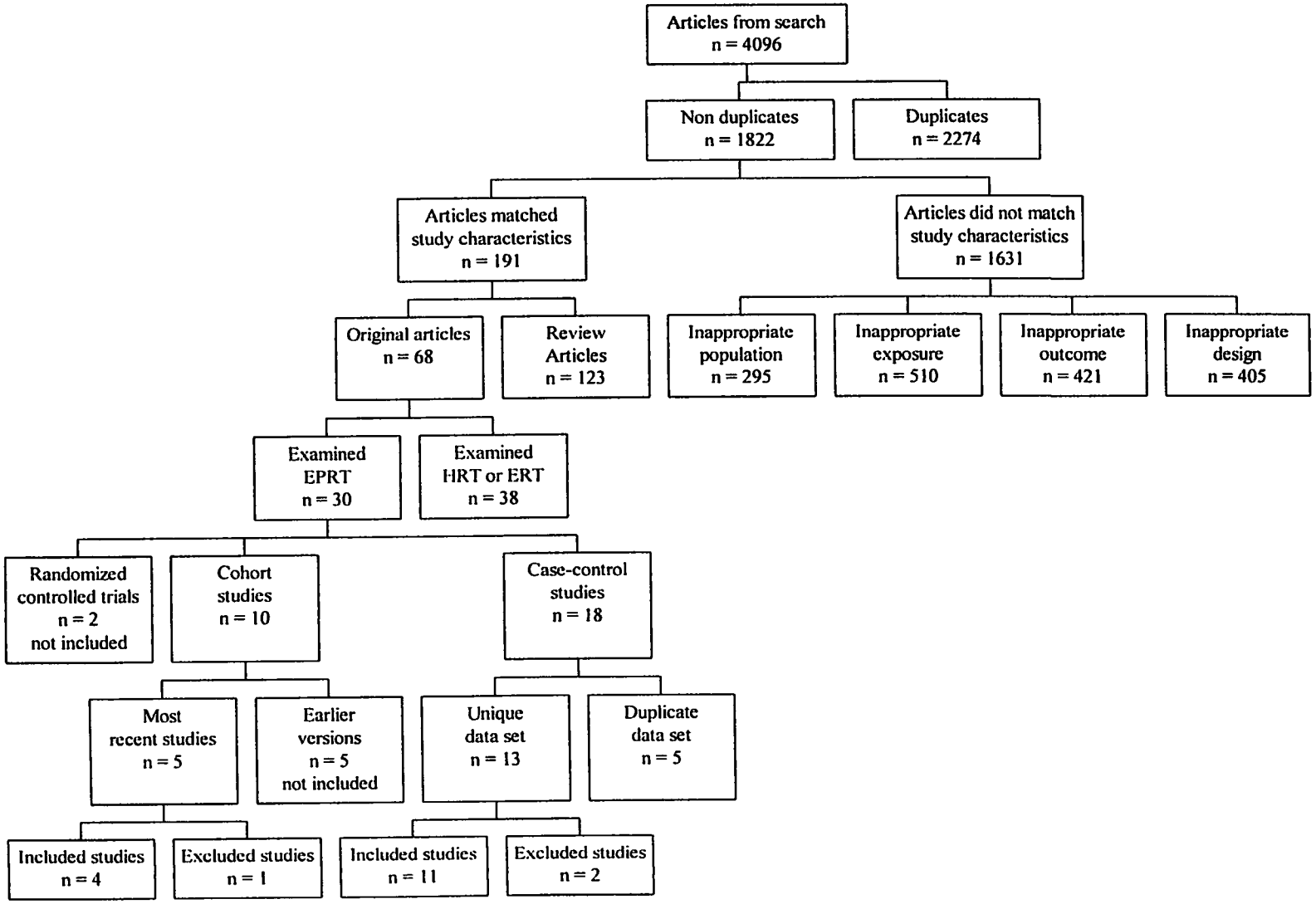
The search strategy revealed a total of 4,094 articles. Of these citations, 1,820 (44.5%) were unique and the remainder (n = 2,274) were duplicates. The results from this search strategy revealed considerable overlap between the databases. Approximately 48% of the articles from Medline were also cited by Cancerlit. There was a 37% overlap between Medline and Current Contents. Medline and Biological Abstracts shared 29% of the citations provided by this search. Similarly, 29% of the citations revealed through

the Embase search were also cited by Medline. Some of the references were cited by all of these databases. Figure 4.1 provides a summary of the screening process involved in the selection of articles included in the systematic review. Of the remaining non-duplicated articles, 1,631 articles (90%) did not meet the inclusion/exclusion criteria.

Most of these articles (n = 510; 31%) were excluded because of an inappropriate exposure. The exposure of interest in this study was a postmenopausal HRT that was composed of estrogen *and* progestin. Many of these excluded studies (n = 305; 60%) examined the use of selective estrogen receptor modulators (SERMs) such as Tamoxifen and Raloxifene. SERMs have an anti-estrogenic effect on breast tissue but an estrogen-like effect on bones[105]. Reasons for their use include reducing the incidence of breast cancer (Tamoxifen) or reduce the risk of osteoporosis (Raloxifene) among women at high risk for the disease. Other studies (n = 96; 19%) looked at exposure to hormonally active agents such as DDT that have estrogenic or androgenic properties. Some of the studies (n = 45; 9%) examined exposure to oral contraceptives, which are also made up of estrogens and progestins. A number of studies looked at endogenous levels of estrogen and progestin (n = 30; 6%). A small number of studies (n = 34; 6%) dealt with exposures relating to alcohol and smoking.

Many articles (n = 421; 26%) were excluded because of an inappropriate outcome. The outcome of interest in this study was incidence of invasive breast cancer. Some of these studies (n = 24; 6%) looked at mortality from breast cancer. Others studied breast cancer in terms of its histology, tumour type, and disease aggressiveness (n = 14; 3%). Others focused on changes in breast density (n = 48; 11%) or benign breast disease (n = 5; 1%). Prevention, prognosis and survivorship from breast cancer were

Figure 4.1 Flow chart describing the selection process for articles included in the systematic review and meta-analysis examining the relationship between breast cancer risk and EPRT exposure among post-menopausal women.



other outcomes that were studied (n = 28; 7%). However, most of the studies that were rejected had outcomes that did not relate at all to breast cancer. These included studies on endometrial, ovarian or colon cancer (n = 116; 28%); bone density and osteoporosis (n = 85; 20%); stroke and heart disease (n = 80; 19%) and a collection of other illnesses (renal disease, thyroid condition, immune function; n = 21; 5%).

Approximately 25% (n = 405) of the articles identified in the search were rejected because of inappropriate design. Many were review articles (n = 298; 74%) that provided information for menopausal women on the risks and benefits of HRT. Other studies (n = 86; 21%) provided details on the pharmacokinetics of ERT or EPRT. The remaining articles rejected in this category (n = 21; 5%) concerned genetic modeling and the development of a breast cancer risk model to help discriminate women at a higher risk of breast cancer.

The remaining 18% (n = 295) of the articles were excluded because of an inappropriate population under study. The population of interest for this study was postmenopausal women with no prior history of breast cancer. Most of these studies (n = 150; 51%) were rejected because the women involved in the study had a prior history of breast cancer. Some of the studies (n = 11; 4%) reported on pre-menopausal women only and other studies (n = 8; 3%) examined only diabetic women. A large number of studies examined the effects of estrogen and/or progestin at the cellular level (n = 75; 25%) or in non-human (eg. mice, primates) populations (n = 51; 17%).

4.1.2 Articles matching study characteristics

There were 191 articles that matched the study characteristics. Of these, 65% (n = 123) were review articles. These reviews were not included but were retained for future reference. In addition, the references cited in these reviews were examined to see if there were any additional studies that were missed in the original search.

The remaining 68 articles were retrieved and each paper was carefully reviewed to see if the study collected data on the relationship between exposure to EPRT and incidence of invasive breast cancer. Of these studies, 42% (n = 30) examined the relationship between exposure to EPRT and breast cancer. The rejected articles (n = 38) examined either ERT or the more general category, HRT. While the articles on HRT could have technically included exposure to EPRT, the authors did not provide any specific details about EPRT use. It is important to note that the majority of these studies examined HRT use in the 1980's, when EPRT use was relatively uncommon. A notable exception is the recent study by Olsson et al[106]. While they did not analyze the brand or type of HRT used, they did state that the predominant formulation currently prescribed in Sweden is EPRT, except among women who have had a hysterectomy. However, the risk estimates that they provided did not distinguish between EPRT and ERT use and thus the data could not be included in this analysis.

4.2 Selected studies

There were a total of 30 studies (18 case control studies, 10 cohort studies and two RCTs) that were obtained from all of the searching sources (electronic databases, review of literature cited in retrieved articles, handsearching and advice from experts). The

characteristics and concerns surrounding these studies are described in the following sections.

4.2.1 Case control studies

There were 18 case control studies that satisfied the inclusion/exclusion criteria outlined in Table 3.1 (Section 3.2). After retrieving these studies and carefully reviewing their methods and results, it was evident that five of these studies reported follow-up, identical to similar data sets that had been published in earlier studies. For instance, the data reported in Pike and Ross[107] in the journal Steroids also appeared in an article published by Ross et al[40] in the Journal of the National Cancer Institute. Franceschi [108] published an Italian translation of a British Journal of Cancer article by La Vecchia et al [109] in the journal Annali dell Instituto Superiori di Sanita. Newcomb et al [110] updated the original Newcomb et al[111] article. Information from the original article was used for valuable background information while the risk estimates were extracted from the most recent article. Ewertz[112] was updated in Ewertz[113] and the most recent article was used. Li et al[114], published in the journal Cancer, was a follow-up of the study by Stanford et al[115], published in the Journal of the American Medical Association. While both publications examined the incidence of breast cancer among the same population of women, Li and colleagues (including Stanford) included an analysis on different histologic types of breast cancer. Since the primary outcome of interest for this systematic review and meta-analysis was invasive breast cancer, the data published in the earlier study was more pertinent and used for the meta-analysis.

The information abstracted from the 13 unique case control studies is summarized in Table 4.1. Date of publication ranged from 1987 to 2000. Eight of the studies (62%) were conducted in North America, the remaining five took place in Europe. All of these studies were published in English. Several of the early studies (pre 1995) relied on hospital admissions to identify breast cancer cases; the more recent studies relied primarily on cancer registries or clinics. The controls were recruited from a number of venues, including voter lists, community registries or health care centres. Two of the studies[116;117] were unpublished theses.

The number of breast cancer cases involving women on EPRT varied considerably among the studies and increased over time. For studies published before 1995, the number of cases ranged from four to 28 (mean 19). For studies published after 1995, the number of cases ranged from 29 to 221 (mean 77.3). A notable exception is the study by Ewertz[113], published in 1993, which reported on 111 cases of invasive breast cancer among women on EPRT.

In most studies, there was one control that had been recruited for every case; i.e. 1:1 matching. A notable exception was the study by Palmer et al[118] and Persson et al[119]. In the Palmer study, there were two controls for every case. Both controls came from the community, identified from a tax assessment list of all Ontario residents. There was a 4:1 ratio in the Persson study. In this nested case control study, cases and controls were made up of women involved in a population-based mammography screening program. The controls were recruited from the women not diagnosed with breast cancer.

In most studies, controls were recruited from the community. There were two studies[109;120] that recruited controls from the same hospitals where the cases had been

Table 4.1 Characteristics of case control studies examining the association between breast cancer risk and exposure to EPRT among postmenopausal women

Study author(s), Year	Location of study	Source of Cases	Number of cases		Source of Controls	Number of controls		Odds ratio (95% CI)	Variables controlled in the analysis
			EPRT	No HRT		EPRT	NoHRT		
Ross et al 2000*	California, USA	Cancer registry	425	873	Local community street search	324	784	1.53 (no CI)	Age at menarche, parity, age at natural menopause, family history of BC, history of BBD, weight, alcohol use, OC use
Chen 1999	Washington state, USA	Health care clinic	221	243	Health care clinic	190	271	1.51 (1.09-2.08)	Age; year of diagnosis Number of mammograms
Kirsh 1999	Ontario Canada	Ontario Cancer Registry	48	272	Ontario Property Assessment database	33	283	1.23 (0.72-2.11)	Age; age at menopause; type of menopause; history of benign breast disease
Magnussen et al 1999	Sweden	Cancer registry	409	1,738	National Population registry	295	2,201	1.63 (1.37-1.94)	Parity; age at first birth; age at menopause; type of menopause; BMI; height
Newcomb et al 1998, 1995	Wisconsin, Mass., Maine, New Hampshire	Cancer registry	138	2,207	Licensed drivers; Medicare users	157	2,574	1.44 (1.19-1.75)	Age at menarche; age at 1st pregnancy; type of menopause, age at menopause; history of BBD & BC; alcohol consumption and education
Persson et al 1997	Sweden	Screening Mammography Program	29	317	Screening Mammography program	79	1,293	1.4 (0.9-2.2)	Age at menarche; parity; age at 1st birth; family history; BMI

* These studies were excluded from the meta-analysis because of methodological flaws

Table 4.1 Characteristics of case control studies examining the association between breast cancer risk and exposure to EPRT among postmenopausal women cont'd.

Study author(s) Year	Location of study	Source of cases	Number of cases		Source of controls	Number of controls		Odds ratio (95% CI)	Variables controlled in the analysis
			EPRT	No HRT		EPRT	No HRT		
La Vecchia et al 1995	Italy	Hospital admissions	6	2,376	Hospital admissions	4	2,395	1.6 (0.4-6.3)	Age; parity; BMI; location, education
Stanford et al 1995	USA	Cancer registry	58	82	Community, random dialing	86	187	0.9 (0.7-1.3	Age, age at first birth; age at menopause; family history of BC
Ewertz 1993	Denmark	Cancer registry	111	997	National Population registry	71	945	1.36 (0.98-1.87)	Age at diagnosis; menopausal status; place of residence
Yang et al 1992	B.C., Canada	Cancer registry	22	430	Community, Voter List	19	432	1.2 (0.6-2.2)	Age; age at menarche; parity; age at first birth; body weight; family history of breast cancer; history of breast biopsy
Kaufman et al 1991	USA and Canada	Hospital admissions	22	1,293	Hospital Admissions	16	1,620	1.7 (0.9-3.6)	Age; age at menarche and menopause; type of menopause; BMI; race; religion; OC use; alcohol
Palmer et al 1991	Toronto, Canada	Cancer hospital	4	493	Ontario tax assessment roll	14	991	0.6 (0.2-2.0)	Age at menopause; type of menopause; BMI
Hunt et al 1987*	Britain	Menopause clinics	28	25	Same as cases	42	55	1.47 (0.75-2.88) unadjusted	None

*These studies were excluded from the meta-analysis because of methodological flaws.

recruited. In the Kaufman study[120], controls were cancer and non-cancer patients while La Vecchia et al[109] selected only non-cancer patients. The effects of including and excluding studies that used hospital controls were examined in the sensitivity analysis (section 4.2.3).

There were a variety of breast cancer risk factors[55] that were controlled in case control studies. These included age (all studies); age at menopause (eight studies); body mass index (BMI; weight/height^2; seven studies); type of menopause (six studies); family history of breast cancer (five studies); age at menarche (five studies); parity (five studies); age at first birth (five studies) and alcohol consumption (three studies).

Age is an important risk factor for breast cancer[55]. When recruiting the controls, most investigators attempted to match the age of the control with the age of the case (except Kaufman and La Vecchia studies), thereby controlling a major risk factor within the design of the study. The age of most controls matched the cases within one to five years. However Palmer et al[118] matched ages by decade. A few studies also matched on geographic location[109;118].

Most studies controlled for the risk factors of breast cancer in the study analysis. The ORs, expressing the breast cancer risk associated with exposure to EPRT, were adjusted for a variety of risk factors (see Table 4.1). The majority of studies (n = 8) controlled for age at menopause. The older a woman is when she begins menopause, the higher her risk of breast cancer[55]. Women have an increased risk of breast cancer if they have a close relative (mother, sister, daughter) who had breast cancer, began menarche at a younger age, are nulliparous or if they do have children, had them after the age of 35.

Five case control studies controlled for family history of breast cancer, age at menarche, parity and age at first birth. There is also some evidence that alcohol consumption increases breast cancer risk[121]; three studies controlled for this factor.

The ORs which described the breast cancer risk associated with "ever" use of EPRT therapy are presented in Table 4.1. Exposure to EPRT was assessed through questionnaires (n = 6), interviews (n = 6) or through computerized pharmacy records. OR values ranged from 0.6 to 1.7, with 11 studies reporting an OR greater than 1.0. Three recent studies[74;110;116] all reported a significant association between ever use of EPRT and breast cancer. There were two case control studies[40;122] that were excluded at this point of the analysis, as described below.

The Hunt study[122]was designed to examine the effects of HRT on the cancer incidence among British women. Recruitment of women from menopausal clinics occurred primarily between 1974 and 1981. Forty-three percent of the women in the study (n = 1954) had been exposed to EPRT, but the authors claimed that the dosage and duration of the progestin component was considerably lower than that found in "modern" formulations. Therefore, it may not be appropriate to include this study with more recent studies because of different formulations. Unfortunately, they failed to record the number of days progestins were prescribed in each cycle and the dose of progestins used. Thus the relative strength of the progestin formulation could not be assessed.

Cases also had lower parity and their first birth occurred at a more advanced age than the controls. However, parity and age at first birth, known risk factors for breast cancer, were not controlled in the design of the study. Because of these differences between cases and controls, it would be important to adjust for these factors during the

analysis of the data when calculating the OR. Unfortunately, the authors only provided adjusted OR for all HRT use (OR = 1.59; 95% CI 1.18-2.10) and did not calculate a separate OR for EPRT use. While an unadjusted OR was calculated from the data provided (OR = 1.47; 95% CI 0.75-2.88), it would not be appropriate to include it in the meta-analysis because it failed to adjust for these risk factors within the sample. All other ORs used in the meta-analysis were adjusted for known risk factors.

Ross et al[40] was also omitted from the meta-analysis. They expressed their results as the breast cancer risk per five years of EPRT use and claimed that there was an increased risk of breast cancer associated with EPRT use. The women in this study averaged only 43 months (or 3.5 years) on EPRT. Therefore the validity of their extrapolation is questionable, particularly since it is still not clear whether or not a linear relationship exists between duration of EPRT use and breast cancer risk. Archer et al[123] raised similar concerns about these results.

Unadjusted ORs with corresponding 95% CI were calculated using the number of cases and controls exposed and not exposed to EPRT that were reported in the Ross et al study. None of these ORs indicated a significant association, even though the authors claimed that one existed. Several attempts to contact each of the authors about this discrepancy were made but efforts to gain access to the standard errors were unsuccessful. In a published response to their non-reporting of confidence intervals, Ross et al[124] stated that "providing confidence intervals around every estimate of risk is not a sensible approach". Because no confidence intervals were reported, this data could not be entered into the calculations of the summary estimates.

Other concerns have also been raised about the Ross study. As was the case with the Hunt study, it would not be appropriate to combine the unadjusted OR that were calculated with adjusted ones obtained from the other case control studies. Therefore, Ross et al was also excluded from the meta-analysis.

The ORs from 11 remaining case control studies were entered into the meta-analysis. The results appear in section 4.4. The effect of including and excluding the Hunt and Ross studies was examined in the sensitivity analysis (section 4.2.3).

4.2.2 Cohort studies

There were 10 cohort studies that satisfied the inclusion/exclusion criteria. Eight of these represented either the original or subsequent follow-up studies. There has been no follow-up for two of the studies: Lauritzen and Meier[36]and Gambrell[37] since they were originally published in 1984 and 1986, respectively. The remaining eight studies were based on data collected from three long term cohorts (Nurse's Health Study, a Swedish cohort, Breast Cancer Detection Demonstration Project). Three of the papers [30;71;125] report on participants in the Nurse's Health study, an American cohort established in 1976. Information for this review relied on information published in the two most recent publications[71;125]. There were three additional papers that followed a Swedish cohort, initially recruited from 1977-1980[126]. A 13 year follow-up was reported in Persson et al[127]. A report on a sub-cohort appears in Persson et al[38]. The latter two reports were used in the meta-analysis. The two remaining papers report on another American based cohort, the Breast Cancer Detection Demonstration Project. The original study was conducted between 1973 and 1980 and the first follow-up (up to 1989) was published in

1994 by Schairer et al[128]. A second follow-up of the study (up to 1995) was published in Schairer et al[39]. The information used in the meta-analysis came from the most recent publication.

The main characteristics of these five cohort studies are summarized in Table 4.2. Three of the cohorts were from the United States and the remaining two were from Europe. The number of women on EPRT varied from 122 in the Lauritzen and Meier study to over 5,500 in the Persson study. It was difficult to discern the total number of women on EPRT in the Colditz and Schairer studies, but both studies reported over 100 cases of breast cancer among EPRT users. Since adjusted RR values and their 95% CI were provided in each of these two studies, the actual number of cases was not needed for the analysis.

Follow-up times varied from five years in the Lauritzen and Meier study to 16 years in the Colditz study. The patient-years of observation for women on EPRT also varied considerably among studies. The smallest value was 521 patient-years in the Lauritzen and Meier study, reflecting a very small population of women on EPRT and a relatively short mean duration of follow-up. In the remaining cohort studies, patient-years of observation ranged from 16,466 patient-years in the Gambrell study to 74,957 patient-years in the Persson study.

Four of the five cohorts provided adjusted estimates for the RR of breast cancer associated with EPRT use. The estimates were adjusted for age and three of the four were adjusted for body weight or BMI, age at menopause and type of menopause. Two of the estimates were also adjusted for parity, age at menarche and age at first birth. It is

Table 4.2 Characteristics of cohort studies examining the association between breast cancer risk and exposure to EPRT among postmenopausal women

Study authors, Year	Location of study	Years of follow up	Patient years of observation	Total EPRT exposure	Source of cohort	Number of cases		Relative risk (95% CI)	Variables controlled in the analysis
						EPRT	NoHRT		
Lauritzen and Meier 1984	Germany	1968-1983	521	4.2 years	Two private practices	4	5	0.96 (0.54-1.53)	Age, type of menopause, body weight, parity, social status
Gambrell 1986*	United States	1975-1984	16,466	2.7 ± 2.4 years	Military hospital	11	22	0.19 (0.02-1.85) unadjusted	None
Colditz et al 1995	United States	1976-1992	28,946	N/A	Nurse's Health Study	111	923	1.41 (1.15-1.74)	Age; type of menopause; parity; age at menarche, first delivery and menopause; family history of BC; history of BBD
Persson et al 1996	Sweden	1977-1991	74,957	N/A	Uppsala Health Care Region Residents	72	48	1.5 (1.2-1.8)	Age; type of menopause, BMI; age at menarche and first delivery and menopause; education; follow up time,
Schairer et al 2000	United States	1973-1995	17,428	3.6 years	Breast cancer detection project	101	761	1.3 (1.0-1.6)	Age; BMI; age at menopause; mammographic screening

* These studies were excluded from the meta-analysis because of methodological flaws.

important to note that Lauritzen and Meier adjusted their variables in the design of the study, while the other three (Colditz, Persson and Schairer) adjusted in the analysis.

Gambrell[37] did not provide any details about the sample of women in his study on EPRT. There was no mention of how many women were on EPRT and for what duration, except to say that there were nine current EPRT users, two past users and 11 patients diagnosed with breast cancer. Unlike all of the other studies included in this analysis, Gambrell did not control for age or any other possible confounding factors. Therefore, there was no way of determining if the women on EPRT differed from the non users in terms of their age, parity, age at menopause and other risk factors. There was also no method to control for these and other risk factors. It was also unclear how these women were recruited into the study.

Gambrell did not provide an estimate of the risk associated with EPRT use. An unadjusted RR was calculated from the incidence data provided and the 95% CI were calculated using the test-based method outlined in Hennekens and Buring[129]. However, as was the case with the Hunt and Ross case control studies, it would not be appropriate to combine this unadjusted RR ratio with the adjusted ones provided by the other cohort studies. Therefore, to maintain consistency in the choice of studies entered into the meta-analysis, the Gambrell study was excluded. The effect of including and excluding this study from the analysis was examined in the sensitivity analysis (section 4.2.3).

4.2.3 Randomized controlled trials

The search strategy revealed two RCTs: Nachtigall et al[130] and Hulley et al[19]. While both involved postmenopausal women on EPRT, they failed to meet the

inclusion/exclusion criteria and were thus excluded from the meta-analysis for reasons described below.

Nachtigall et al[130] was a 22 year study involving 168 postmenopausal women who were hospitalized for the entire study period for a variety of chronic conditions that, according to the authors, were unrelated to breast cancer. For the first 10 years of the study, half of the women were randomly assigned to the treatment group and received EPRT. EPRT consisted of daily doses of conjugated estrogens (2.5 mg/day) and a seven day cycle of medroxyprogesterone acetate (10 mg/day) each month. The other 84 women received a placebo. Women were matched on age, medical diagnosis for which they were hospitalized (eg. diabetes, multiple sclerosis, rheumatoid arthritis) and history of cigarette smoking. After the initial 10 years, researchers reported four cases of breast cancer among the women in the placebo group and no cases of breast cancer among the women on EPRT. They provided no information about how breast cancer was ascertained. At this point in the study, the original group designation was revealed. After receiving this information, the participants were then given the option of remaining with their original group or switching. Thus, women from the treatment group could remain on EPRT or discontinue the therapy. Women in the placebo group could remain off the therapy or begin taking EPRT.

For the next 12 years of the study, there were 45 women who remained on EPRT and 32 women from the original placebo group who chose to begin EPRT. There were 42 women who remained off EPRT and 33 women who chose to discontinue use of EPRT. Therefore, the second phase of the study involved 77 women on EPRT and 75 women taking no HRT. A total of six women in the original treatment group and 10 in

the placebo group were lost to follow-up or died. The results from this study are described in Table 4.3.

Table 4.3 Results from the Nachtigall et al RCT

	Breast cancer	**No Breast Cancer**	**Total**
EPRT Current use for 22 yrs	0	45	45
EPRT Current use for 12 yrs	0	32	32
EPRT Former use for 10 yrs	0	33	33
Never users of EPRT	6	36	42
Total	6	146	152

The authors of this study conclude that there was no increased risk of breast cancer associated with EPRT use because none of the women that had been exposed to EPRT developed breast cancer. They argued that progesterone may protect against breast cancer.

To date, this is the longest study examining the relationship between EPRT use and breast cancer and one of the few RCTs. While these results appear to support the notion that there is no relationship between EPRT use and breast cancer, there are a number of flaws in this study that lead one to question the validity of these results and their applicability to the general population of postmenopausal women.

All of the women involved in this study were long term patients in a chronic care hospital. They were being treated for a variety of conditions that appear to be unrelated to breast cancer (heart disease, epilepsy, Parkinson disease). In Phase I of the study, the women were matched on medical diagnosis and this should control for disease

confounding. They were also matched on age and smoking status, two risk factors associated with breast cancer. The authors also stated that they were randomly assigned to the treatment/control group, but there was no description as to how this was accomplished. In the second phase of the study, the allocation of the women to placebo and treastment groups was not random.

Parity is an important risk factor that is frequently controlled in studies examining the association between exposure to HRT and breast cancer. In a recent review on the epidemiology of breast cancer, Key and colleagues[55] report that nulliparous women have a 25% increased risk of breast cancer compared to women who have had at least one full term pregnancy. Therefore, nulliparity places women at a higher risk for breast cancer. However, women in this study were not matched on parity. There were 41 nulliparous women (49%) in the control group and only 22 nulliparous women (26%) in the treatment group. Therefore, the higher incidence of nulliparity in this group rather than their lack of EPRT exposure could explain the higher incidence of breast cancer among women in the control group.

The controlled setting of a hospital would have ensured that the women in the treatment group would be taking EPRT as prescribed. However, all of the women in this study would be less physically active than the general population and this would serve to increase their risk of breast cancer[55]. They are likely exposed to more medications than average due to their chronic conditions. Therefore the applicability of these findings to the general population is questionable.

Phase II of the study permitted women to chose which treatment they received. This could have introduced further bias into the study because "healthier" women may

have chosen to remain on or switch onto the therapy and "less-healthy" women may have chosen to remain off the therapy or discontinue use. At this point, the study was no longer a RCT and interpretation of the findings become more difficult.

The Nachtigall study was thus excluded because the general applicability of the findings were questionable as the participants were long term (20+ years) hospital patients. In addition, almost 50% of the placebo group were nulliparous (compared to 26% of the treatment group) and therefore, in terms of parity, had a higher risk profile for breast cancer than the treatment group

Hulley et al[19] was a randomized, blinded, placebo controlled secondary prevention trial that was part of the Heart and Estrogen/progestin Replacement Study (HERS). While the primary outcome was the occurrence of myocardial infarction or death due to coronary heart disease, the authors also looked at secondary outcomes such as breast cancer. They found no significant difference in the incidence of breast cancer among women in the treatment and control groups (RR = 1.30; 95 % CI 0.77-2.19). There were 32 women (2.3%) on EPRT and 25 women (1.8%) in the placebo group that developed breast cancer during the four year study. This finding suggests that there is no increased risk of breast cancer associated with exposure to EPRT. However, there are some problems with this study that would call into question the validity of this conclusion.

The length of follow-up in this study averaged 4.1 years. While it is difficult to pinpoint an exact latency period for breast cancer, most of the estimates start at five years and can reach upwards to 30 years for exposures such as environmental pollutants[131]. Most studies that have examined the relationship between breast cancer and HRT use

have follow-up periods for at least five years or longer. A longer follow-up in this RCT may reveal an association between breast cancer and EPRT use.

There was an additional problem concerned exposure to EPRT. Studies that examine the association between HRT and breast cancer traditionally compare the incidence of breast cancer among two groups: ever users (those exposed to HRT; the treatment or exposed group) and never users (those never exposed to HRT; the placebo or control group). In this study, there were 318 women (23%) in the placebo group who had been prior users of ERT. Therefore, the "never user" group did not exist. Because of this, it may not be reasonable to combine the results of this study with the other studies that make a clearer distinction between ever and never use.

A similar problem existed in the treatment group. The aim of the HERS Trial was to examine the consequences of exposure to EPRT. While all of the women in the treatment group (n = 1380) were exposed to EPRT, there were 331 women (24%) who also had prior exposure to estrogen while on ERT. In this RCT, details about the duration or formulation of the estrogen that the women had been exposed to prior to the RCT were lacking. However, the average age of women in the study was 66.7 years, so theoretically there could have been women on ERT for 10 years or more. In the case control and cohort studies examining the relationship between breast cancer and EPRT, women with prior ERT exposure were excluded from the analysis. Therefore, the results of this RCT would not be comparable to the other studies that have been completed. Therefore, calculating the RR of breast cancer associated with EPRT use in this study has several associated flaws.

For the reasons outlined above, the results from both RCTs were not introduced into the meta-analysis. The effect of including and excluding one or both of these RCTs was examined in the sensitivity analysis (section 4.4.3).

4.3 Quality assessment

4.3.1 Quality assessment of case control studies

According to the NOS, the overall quality of case control studies used in the meta-analysis was quite high (Table 4.4). The mean score for all 11 studies was 7.1 ± 0.7 out of a possible maximum of nine. The scores ranged from a top score of eight (n = 3 studies) to the low score of six (n = 2). The remaining studies (n = 6) all had a score of seven. This suggests two things. First of all, there was little variation in the quality of the studies included in the meta-analysis. Secondly, since the maximum NOS score is nine, most of the studies used were both well designed and reported according to the three NOS criteria.

As described in section 2.6.3.1, there were three main criteria that were used to evaluate each study: selection; comparability and exposure. The majority of studies provided an adequate case definition that consisted of histological confirmation through medical or hospital records or cancer registry. Cases in most studies consisted of postmenopausal women, newly diagnosed with primary invasive breast cancer. One exception was Ewertz[113], where three percent of the breast cancer cases were *in situ*. Most studies (n = 10) explicitly stated that they excluded women with a breast cancer history. It was not clear whether this was the case with the Kirsh[117] study. The lower

Table 4.4 Quality assessment of case control studies using the Newcastle-Ottawa Scale: scores and evaluation of criteria

Study author(s), Year Total NOS score[a]	SELECTION: source/validation of cases, selection of controls NOS score[b]	COMPARABILITY: of cases and controls based on factors controlled in design or analysis NOS score[c]	EXPOSURE: method of ascertainment for cases and controls, non response rate NOS score[d]
Palmer et al., 1991 6★	Cases: all from one Toronto hospital, validated Non random sample of women Controls: from tax assessment roll; residents of Ontario 3★	Design: age Analysis: type, age at menopause 2★	Exposure: structured interview for both (not blind to status) Non response rate: 21% cases; 35% controls 1★
Kaufman et al., 1991 7★	Cases: hospital admissions, validated Random sample of women Controls: from hospital 2★	Design: age Analysis: age, age at menarche 2★	Exposure: structured interview by trained nurses for both (blind to status) Non response rate: 4% for both 3★
Yang et al., 1992 7★	Cases: record linkage; Random sample of women Controls: from BC voter's list 3★	Design: age Analysis: age, age at menarche 2★	Exposure: mail in questionnaire for both (blind to status) Non response rate: 32% for both 2★
Ewertz, 1993 7★	Cases: Danish cancer registry, validated Non random sample of women Controls: from community 4★	Design: age Analysis: menopausal status 2★	Exposure: mail in questionnaire for both (blind to status) Non response rate: 12% cases; 22% controls 1★
La Vecchia et al., 1995 6★	Cases: admitted cases from one hospital; validated Non random sample of women Controls: from hospital 2★	Design: age Analysis: age, parity 2★	Exposure: in hospital questionnaire (not blind to status) Non response rate: < 4% for both 2★

a. maximum score is 9
b. maximum score is 4
c. maximum score is 2
d. maximum score is 3

Table 4.4 Quality assessment of case control studies using the Newcastle-Ottawa Scale: scores and evaluation of criteria cont'd

Study, Year Total NOS score[a]	SELECTION: source/validation of cases, selection of controls NOS score[b]	COMPARABILITY: of cases and controls based on factors controlled in design or analysis NOS score[c]	EXPOSURE: method of ascertainment for cases and controls, non response rate NOS score[d]
Stanford et al., 1995 7★	Cases: Cancer registry; validated Non-random sample of women Controls: from community 4★	Design: age Analysis: family history of breast cancer 2★	Exposure: structured interview for both (not blind to status) Non response rate: 19% of cases and 27% of controls 1★
Persson et al., 1997 7★	Cases: Mammography screening; validated Random sample of women Controls: from mammography screening program 3★	Design: age Analysis: parity 2★	Exposure: questionnaire administered by nurses (blind to status) Non response rate: no mention 2★
Newcomb et al., 1998 8★	Cases: Cancer registry; validated Non-random sample of women Controls: from community 3★	Design: age Analysis: family history of breast cancer 2★	Exposure: telephone interview for both Non response rate: 19% of cases and 16% of controls 3★
Chen 1999 8★	Cases: Group health co-op; validated Non-random sample of women Controls:from same health co-op 3★	Design: age Analysis: body mass index 2★	Exposure: computerized pharmacy data Non response rate: 7% of cases and 8% of controls 3★
Kirsh 1999 7★	Cases: Ontario Cancer registry; valid Random sample of women Controls: from Ontario property database 3★	Design: age Analysis: age at menopause 2★	Exposure: questionnaire for both (not blind to status) Non response rate: 13% of cases and 22% of controls 2★
Magnussen et al., 1999 8★	Cases: Cancer registry; validated Random sample of women Controls: from National population registry 4★	Design: age Analysis: parity, age at menopause 2★	Exposure: mail in questionnaire (blind to status) Non response rate: 36% of cases and 32% of controls 2★

a. maximum score is 9
b. maximum score is 4
c. maximum score is 2
d. maximum score is 3

scores in this criterion were mainly a consequence of using a non random sample of women [109;118] or reliance on hospital controls.

There was no variation in the scores awarded for comparability—all studies were assigned a score of two out of a possible maximum of two. The comparability criterion assessed the similarity between cases and controls according to important risk factors for breast cancer. All studies controlled for age in the design of the study by matching the age of the controls within one to nine years of the age of the cases. All studies also controlled for an additional risk factor in the analysis, by adjusting the OR according to recognized risk factors (see Table 4.1 for a complete list of variables controlled in the analysis). While all of the studies controlled for three or more risk factors, the NOS only awarded a maximum score of two.

There was considerable variation between studies for the exposure criterion. Three studies scored either a one or three and the remaining five studies were awarded a score of two. The top scoring studies in this category obtained estimates of EPRT exposure through pharmacy records or structured interviews where the interviewer was blind to the status of the participant. Studies with long follow-ups (greater than five years) and similar response rates for cases and controls also received higher scores in this category.

4.3.2 Quality assessment of cohort studies

According to the NOS, the overall quality of most cohort studies used in the meta-analysis was high. The results from the quality scoring are summarized in Table 4.5. The

Table 4.5 Quality assessment of cohort studies using the Newcastle-Ottawa Scale: scores and evaluation of criteria

Study, Year Total NOS score[a]	SELECTION: representativeness of Exposed cohort; selection of Unexposed cohort; exposure assessment NOS score[b]	COMPARABILITY: of exposed and unexposed cohorts based on factors controlled in design or analysis NOS score[c]	OUTCOME: assessment; length of follow-up; adequacy of follow-up NOS score[d]
Lauritzen and Meier 1984 5★	Exposed: selected group of users: patients of two doctors Unexposed: from same community Exposure assessment: physician files 2★	Design: age, body weight, parity Analysis: none 2★	Outcome assessment: no description Length of follow-up: 5.3 years Adequacy: no statement 1★
Colditz et al., 1995 7★	Exposed: selected group of users: Nurses from Nurses Health study Unexposed: from same community Exposure assessment: questionnaire 2★	Design: age Analysis: age at menopause 2★	Outcome assessment: record linkage Length of follow-up: 15+ years Adequacy: 5% of women unaccounted for 3★
Persson et al., 1996 8★	Exposed: large sample of women from Uppsala, Sweden Unexposed: from same community Exposure assessment: Pharmacy records 4★	Design: age Analysis: 1★	Outcome assessment: record linkage Length of Follow-up: 11+ years Adequacy: 11-16% of the women unaccounted for 3★
Schairer et al., 2000 8★	Exposed: women from a breast cancer screening program (BCDDP) Unexposed: from same community Exposure assessment: questionnaire 3★	Design: age Analysis: age at menopause 2★	Outcome assessment: pathology or self report Length of Follow-up: 10.2 years Adequacy: all women accounted for 3★

a. maximum score is 9
b. maximum score is 4
c. maximum score is 2
d. maximum score is 3

mean score for the four studies was 6.0 ± 1.9 out of a possible maximum of nine. Two of the studies had a score of eight, one had a score of seven and the remaining study had a low score of five. It is interesting to note that the three studies with the highest score were all relatively recent (published between 1995 and 2000), while the study with the lowest NOS score was published in 1984.

There was very little variation in the quality assessment of the top three studies. The Persson[127] and Schairer[39] studies had similar scores for seven of the eight items and there was only a few differences between the quality scoring of these two studies and the Colditz[71] paper. The Colditz paper reported on the Nurses' Health Study and thus the cohort consisted of a "selected group of users". The population of women studied in the other two cohorts were deemed to be more representative of the average postmenopausal women in the community and thus these two studies were awarded a higher score (three instead of two) in the selection category. The top NOS score for a cohort study is nine. This indicates that these three studies, with their scores ranging from seven to eight, were well designed and reported, according to the NOS criteria for cohort studies.

The format of the NOS cohort scale was similar in format to the NOS case control scale. There were three main criteria that were used to evaluate each study: selection, comparability and outcome. The four items in the selection criterion compared the exposed and unexposed cohort, evaluated how exposure was measured and whether or not the population was free of breast cancer at the start of the study. All studies selected the non-exposed cohort from the same population of women as the exposed cohort. It was clear in the three most recent studies that none of the women had breast cancer before the study began. It was difficult to discern whether this was the case in the

Lauritzen and Meier[36] study. One way in which this early study was better than the other three was in its ascertainment of exposure. The more recent studies relied on a written self report of hormone use while this 1984 study used the physician records to measure exposure. Another important consideration is how well the cohort represents the average postmenopausal woman in the community. The Colditz study examined only nurses (but from many different cities throughout the US), while the Lauritzen and Meier study was limited to all the postmenopausal patients (n = 458) in two private practices in Germany. The other two studies involved large groups of women from Sweden[38] or the United States[39].

There was little variation between studies in terms of comparability. Three of the four cohort studies received a maximum score of two because they controlled for at least two breast cancer risk factors. Each of the studies controlled for age in the study design. The Lauritzen and Meier study also controlled for other factors in the study design (Table 4.2). The remaining cohort studies controlled for risk factors in the analysis and presented RRs adjusted accordingly.

Outcome was the third criterion evaluated by the NOS. The three most recent cohort studies all used record linkage to assess breast cancer. The length of follow-up for each study was at least 11 years. It is interesting to note that the actual period of follow-up for these three studies all overlap, encompassing the years 1977 to 1991. The adequacy of follow-up varied somewhat between these three studies, but all fell within the required "less than 20% lost to follow-up". The Lauritzen and Meier study differed considerably from the other three in terms of outcome. First of all there was no description of how outcome was assessed and no statement about the adequacy of follow-

up. It was not clear if all the women were involved throughout the study and there was no mention of women moving or dropping out. The follow-up was considerably shorter in this study; the authors report a mean duration of 4.2 years but do not provide any additional details.

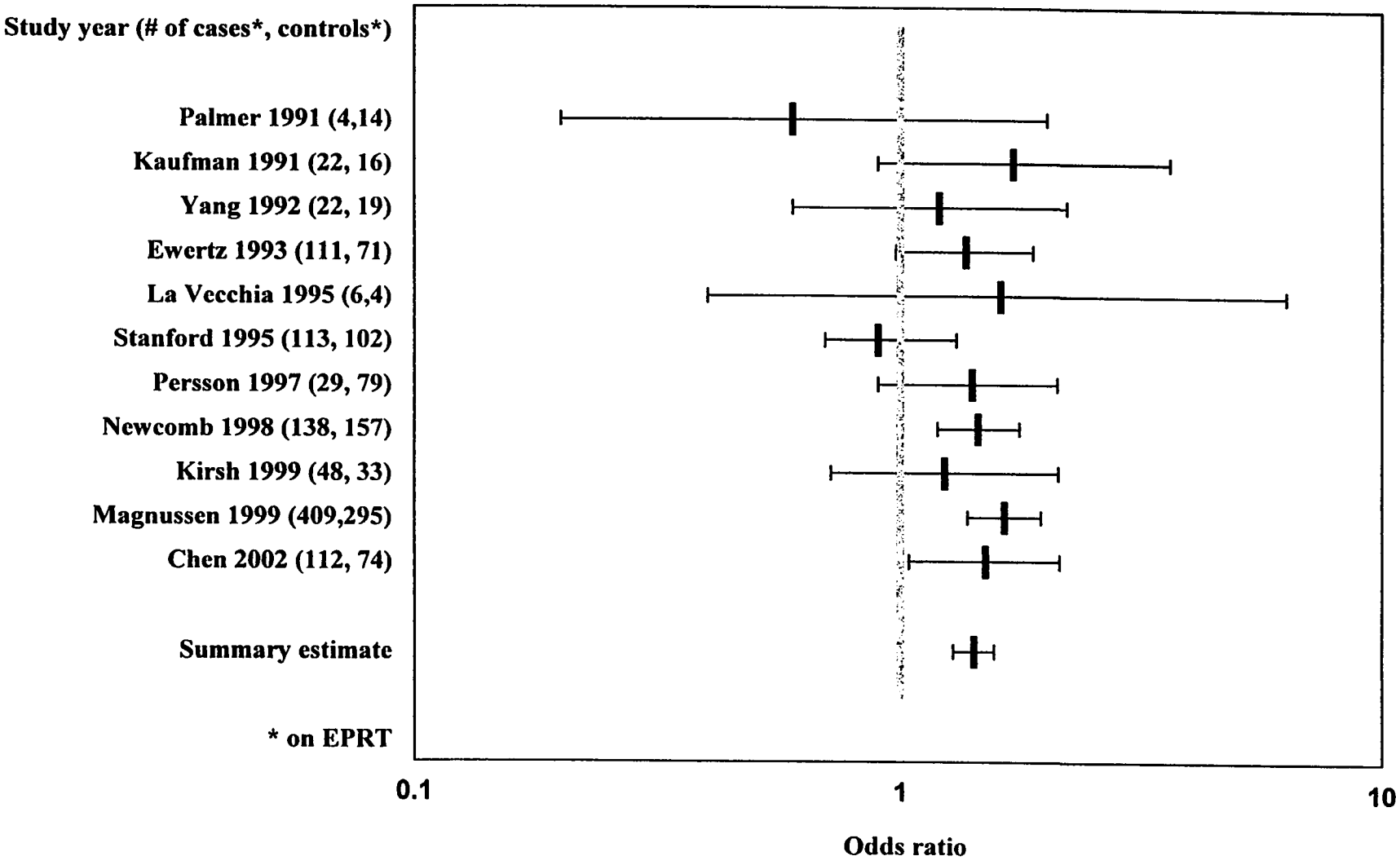
4.4 Quantitative analysis

4.4.1 Ever users of estrogen progestin replacement therapy vs. never users of hormone replacement therapy

4.4.1.1 Case control studies

All of the case control studies compared the risk of breast cancer between ever and never users of EPRT. Ever users of EPRT were defined as women who had indicated that they had used EPRT. Never users were women who had never used any type of HRT. Figure 4.2 plots the ORs (short vertical bar), 95% CI for each case control study (horizontal line) and the summary estimate. The numbers in brackets next to the study author and year indicate the number of cases and controls in the study that were on EPRT. The vertical line serves as a reference to indicate an OR of 1.0, i.e. no risk. ORs located to the right of this line indicate an increased risk of breast cancer associated with EPRT use and values to the left indicate a decreased risk. Among the 11 case control studies, there were nine of the 11 studies that had an OR greater than 1.0, but only three were significant ($p < 0.05$). It is interesting to note that almost all of the confidence intervals for 11 studies displayed in Figure 4.2 overlap and thus the differences between studies are unlikely to be statistically significant.

Figure 4.2 Breast cancer risk with ever use of EPRT vs. never use of HRT among postmenopausal women in case control studies: odds ratios and 95% confidence intervals



The ORs and 95% CI of the breast cancer risk associated with EPRT from each study were combined using the fixed effects model. The summary estimate was 1.42 and the 95% CI was 1.29-1.57. This indicates an increased risk associated with EPRT use among ever users.

Chi-square tests of association and heterogeneity were performed and the results displayed in Table 4.6. The results from these analyses suggest that there is a significant association between breast cancer and EPRT use ($\chi^2_{association} = 48.31$, $p < 0.0001$).

Table 4.6. Results of the chi-square tests of association and homogeneity: ever users of EPRT vs. never users of HRT among postmenopausal women in case control and cohort studies.

	Case control studies	Cohort studies
Summary (95% CI)	1.42 (1.29-1.57)	1.38 (1.23-1.56)
# of studies	N = 11	N = 4
$\chi^2_{association}$	48.30 $p < 0.0001$	27.95 $p < 0.0001$
$\chi^2_{homogeneity}$	12.32 $p = 0.27$	2.80 $p = 0.42$

The Cochran's Q test revealed that the heterogeneity among the 11 case control studies was not statistically significant ($\chi^2_{homogeneity} = 12.31$, $p = 0.27$). Clarke and Oxman[82] report that a p value greater than 0.10 suggests little variation between studies. When the variation between studies is low, they argue that a fixed effects model will provide a reasonable estimate of the confidence interval. However, if the heterogeneity is significant (signalling a large variation between studies), it is more appropriate to use a random effects model which will incorporate the between-study variation in the estimate of effect size. Therefore, the choice of the fixed effects model for this analysis is a

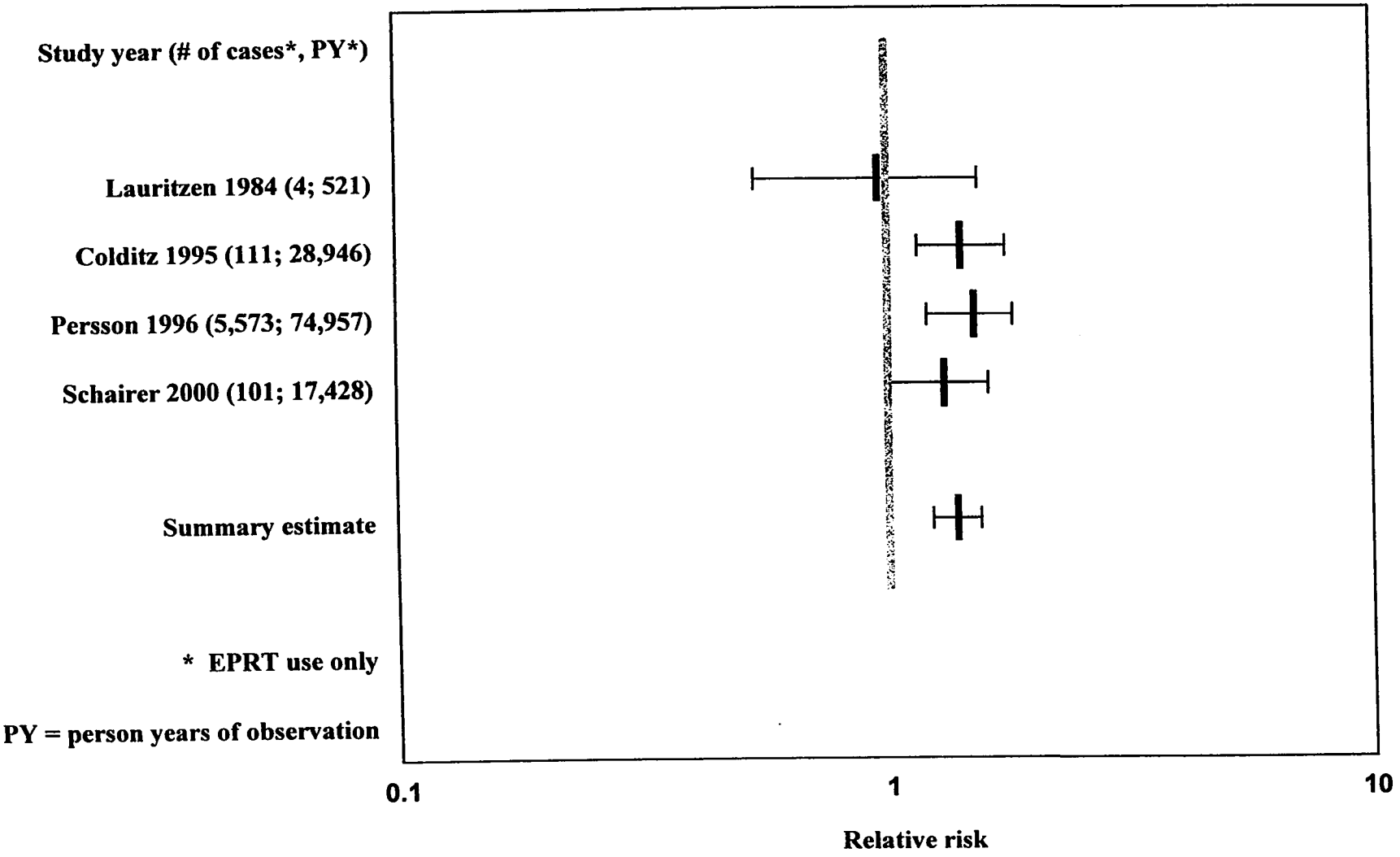
reasonable one and provides an estimate of the average breast cancer risk for those postmenopausal women using EPRT.

4.4.1.2 Cohort studies

The four cohort studies compared the risk of breast cancer between ever and never users of EPRT. Ever users of EPRT were defined as women who had ever used EPRT and never users were women who had never used any type of HRT. Figure 4.3 plots the RRs (short vertical bar), 95% CI for each cohort study (horizontal line) and the summary estimate. The numbers in brackets next to the study author and year indicate the number of breast cancer cases for women on EPRT and the total number of person-years of observation for women on EPRT. The vertical line serves as a reference to indicate a RR of 1.0, i.e. no risk. RRs located to the right of this line indicate an increased risk of breast cancer associated with EPRT use and values to the left indicate a decreased risk. Among the four cohort studies, three had a RR significantly greater than 1.0 ($p < 0.05$). The RRs (and 95% CI) from each study were combined using the fixed effects model. The summary estimate was 1.38 and the 95% CI were 1.23-1.56. This indicates an increased risk associated with EPRT use among ever users of EPRT compared with never users of HRT.

Statistical analyses were also performed on the estimate obtained from the cohort studies. The results from chi-square tests of association and heterogeneity are displayed in Table 4.6. The results obtained from the cohort studies were similar to the case control study results. The results suggest that, based on the results of cohort studies, there is a significant association between breast cancer and EPRT use among cohort studies

Figure 4.3 Breast cancer risk with ever use of EPRT vs. never use of HRT among postmenopausal women in cohort studies: relative risk and 95% confidence intervals



($\chi^2_{association} = 27.95$, $p < 0.0001$). Furthermore, the Cochran's Q test revealed that the heterogeneity among the four cohort studies was not statistically significant ($\chi^2_{homogeneity} = 2.80$, $p = 0.42$). The fact that the p value was greater than 0.10 suggests that there was little between study variation between the four cohort studies. Therefore, the fixed effects model was again a reasonable choice for combining the RRs from the cohort studies to arrive at a summary estimate.

The summary estimate determined from the cohort studies was almost identical to the summary estimate arising from the case control studies (see Table 4.6). This further attests to the robustness of the result and suggests that there is an approximately 40% greater risk for breast cancer for postmenopausal women on EPRT.

Although the heterogeneity between all of case control studies and between all of the cohort studies was not found to be statistically significant, it is evident from Figure 4.2 that there was variation among the case control studies. In an attempt to explain this variation and learn more about the nature of the relationship between breast cancer risk and exposure to EPRT, several subgroup analyses were performed. The results of these subgroup analyses appear in the next section.

4.4.2 Subgroup analyses

4.4.2.1 Study quality

The quality scores of the 11 case control studies, as determined by the Newcastle Ottawa Scale, varied from a low score of six to a high score of eight out of a maximum of nine. The ORs of the 11 case control studies were plotted in order of study quality,

starting with the higher quality studies, to determine if the variation in study quality could explain some of the observed variation between studies (Figure 4.4). When the OR values of each study are displayed in decreasing order of quality, several trends are evident.

First of all, a decline in study quality (from a score of eight to six) tended to be accompanied by a decline in the size of the OR (from 1.6 to 0.6). This suggests that studies judged to be of high quality showed the strongest relationship between breast cancer risk and EPRT exposure, while the weaker studies showed a weaker association between breast cancer and EPRT. Another interesting trend involved the width of the 95% CI. As study quality decreased, the width of the OR's 95% CI tended to increase (from 0.6 to 5.9). This indicates that study quality, as determined by the NOS, is a reflection of the precision of the OR. These results indicate that the higher the study quality, the more precise the estimate.

The range of quality scores for cohort studies was similar to the case control studies. When the RRs were plotted in order of study quality (Figure 4.5), a trend similar to that observed for the case control studies was noted. While a decline in the study quality was associated with a decrease in the size of the RR, the differences between the highest and lowest quality study was not as great (from 0.96 to 1.5). The width of the confidence interval also increased with decreasing study quality, but the difference between the studies with the highest and lowest NOS scores was not as great (0.6 and 0.9, respectively) as that noted with the case control studies.

The ORs of the three case control studies with a NOS score of eight were combined using the fixed effects model in order to obtain a summary estimate. The same

Figure 4.4 Breast cancer risk with ever use of EPRT vs. never use of HRT among postmenopausal women in case control studies, in order of study quality: odds ratios and 95% confidence intervals

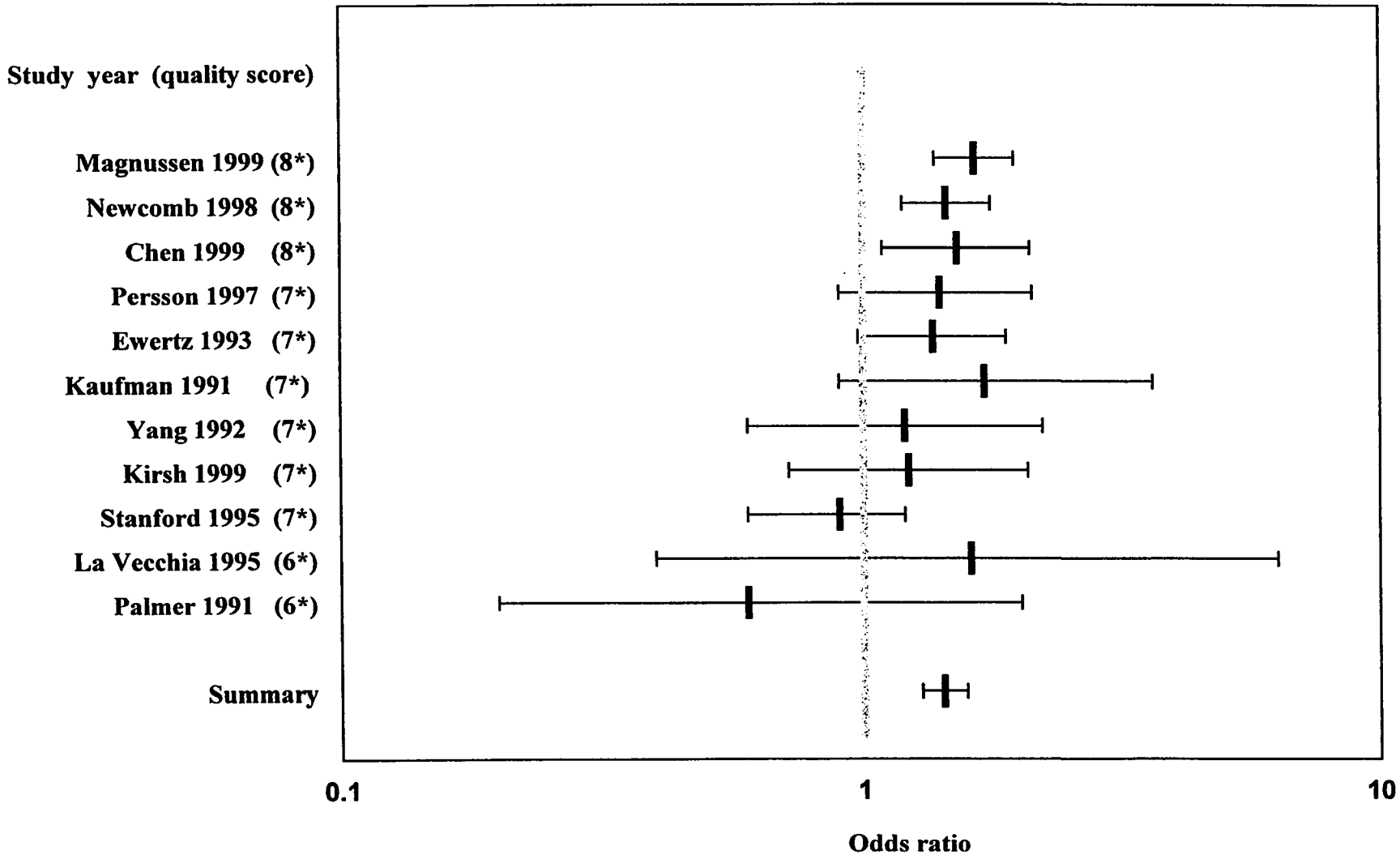
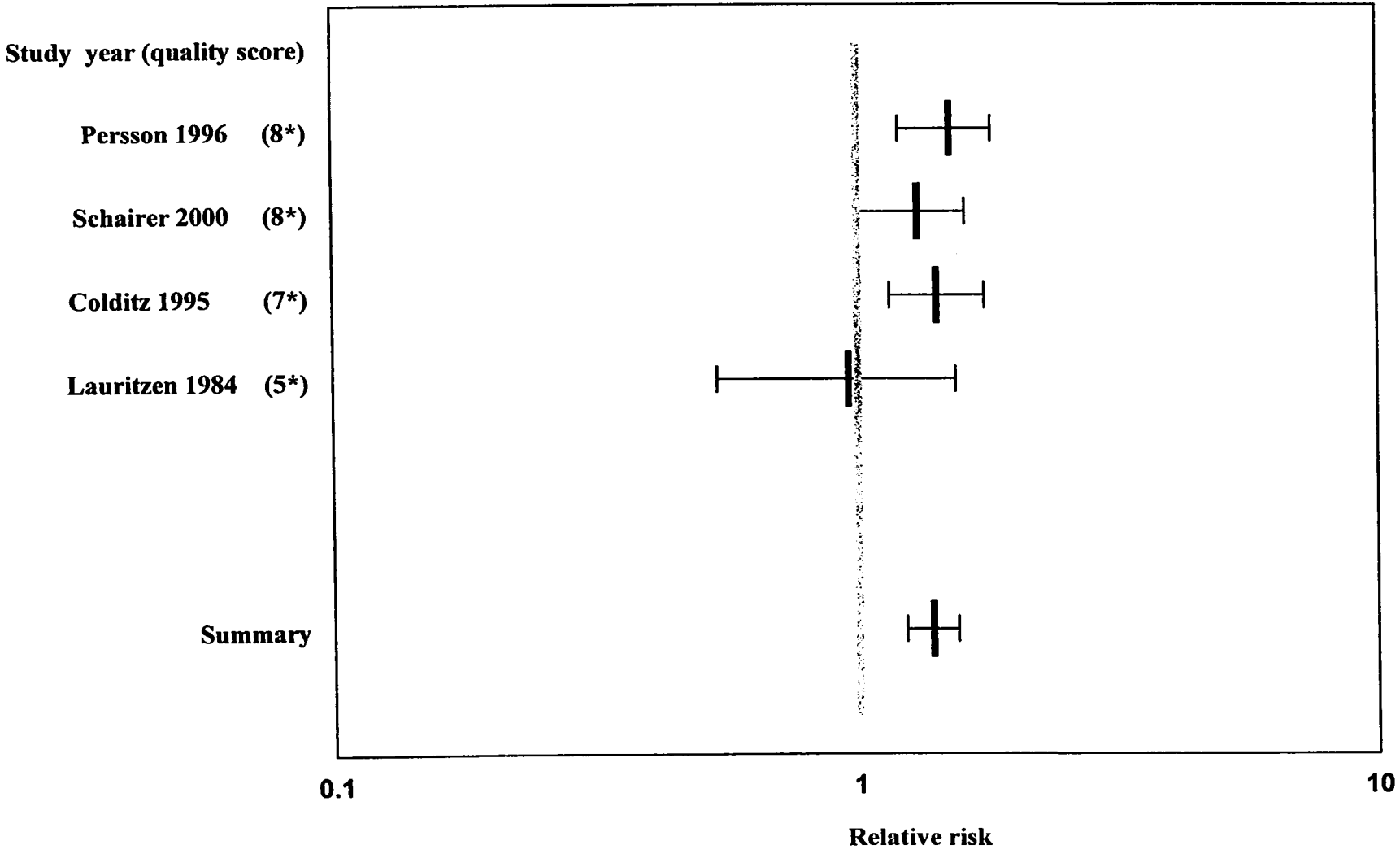


Figure 4.5 Breast cancer risk with ever use of EPRT vs. never use of HRT among postmenopausal women in cohort studies, in order of study quality: relative risk and 95% confidence intervals



process was repeated for the six studies with a NOS score of seven and the two studies with a NOS score of six. The three summary estimates for the three different NOS scores were plotted and the results are displayed in Figure 4.6, along with the overall summary estimate for the 11 case control studies. The same relationships are evident in the plot of the different summary estimates. The highest quality studies had the highest risk estimate and the narrowest 95% CI. As study quality declined, the risk estimate declined and the width of the 95% CI increased.

Chi-square tests of association and heterogeneity were performed for each of the three subgroups. The results of these statistical tests of significance are displayed in Table 4.7. One of the interesting findings that resulted from combining studies with similar NOS scores was a significant breast cancer risk associated with EPRT use among studies with NOS scores of seven and eight but not among studies with a NOS score of six. It was also interesting to note that the association between breast cancer risk and EPRT use was significant among the higher quality studies (NOS score eight or seven) but not with the lower quality studies (NOS score six). Furthermore, there was no significant heterogeneity between the studies with the same NOS score. These results suggest that study quality is an important consideration and differences in the study quality observed among the case control studies could help explain some of the variation observed between studies. When the studies were grouped according to their NOS score, there was no significant heterogeneity observed within each group.

Figure 4.6 Subgroup analysis of study quality: summary estimates and 95% confidence intervals for case control studies with the same quality score.

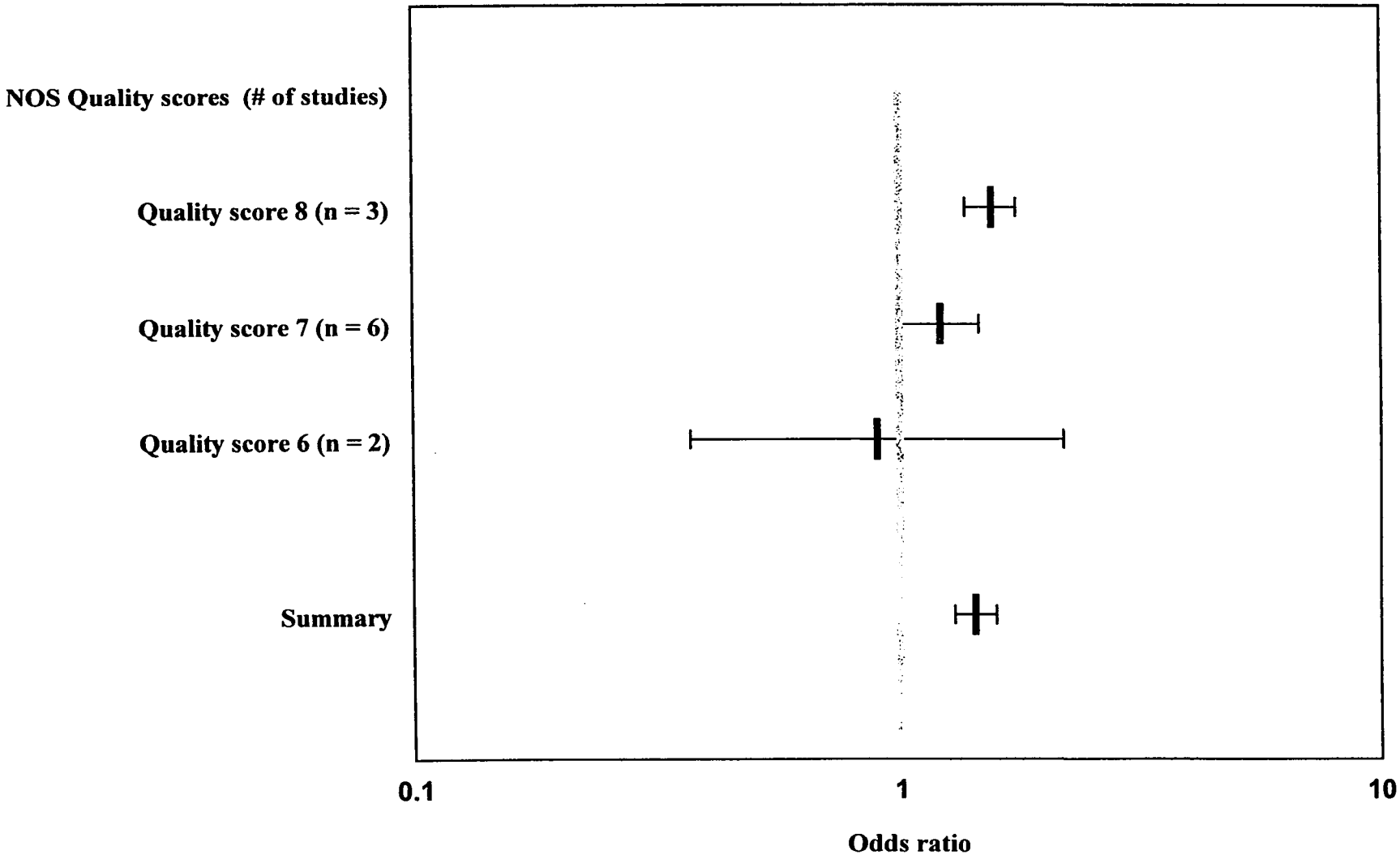


Table 4.7. Results of the chi-square tests of association and homogeneity for subgroup analysis based on study quality: case control studies with the same NOS score.

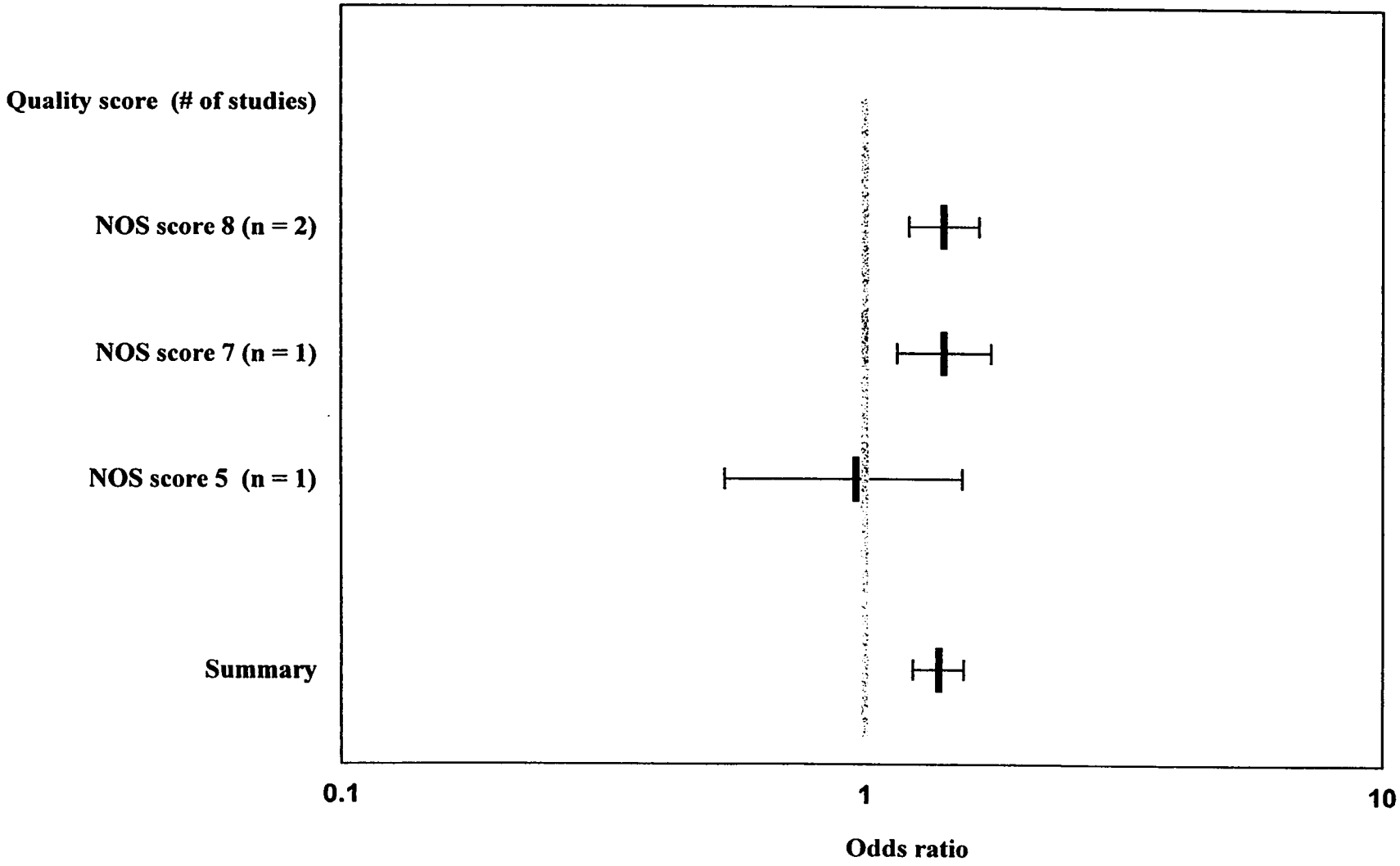
	NOS score 8	NOS score 7	NOS score 6
Summary (95% CI)	1.54 (1.36-1.73)	1.21 (1.01-1.45)	0.90 (0.37-2.17)
# of studies	N = 3	N = 6	N = 2
$X^2_{association}$	49.40 $p < 0.0001$	4.48 $p = 0.03$	0.06 $p = 0.81$
$X^2_{homogeneity}$	0.88 $p = 0.64$	4.64 $p = 0.46$	1.15 $p = 0.29$

There were insufficient studies available to do a similar subgroup analysis for the cohort studies. Figure 4.7 shows the combined estimate for the two cohort studies with a NOS score of eight, plotted alongside the other two studies with NOS scores of seven and five. Unlike Figure 4.6, it is evident in Figure 4.7 that there is little, if any, difference between the RR of the study with a NOS score of seven and the combined RR of the two studies with a NOS score of eight. However, the cohort studies with the higher quality scores report a significant breast cancer risk associated with EPRT use. The cohort study with the lowest quality score found no significant breast cancer risk associated with EPRT use.

4.4.2.2 Duration of estrogen progestin replacement therapy

The case control studies published prior to 1995 did not take into account the duration of time that postmenopausal women had been exposed to EPRT. However, several studies published since 1995 reported on the association between breast cancer and EPRT use based on the number of years that women had been taking EPRT. It was difficult to do a comparative analysis because many of the studies reported ORs based on

Figure 4.7 Subgroup analysis of study quality: summary estimates and 95% confidence intervals for cohort studies with the same quality score.



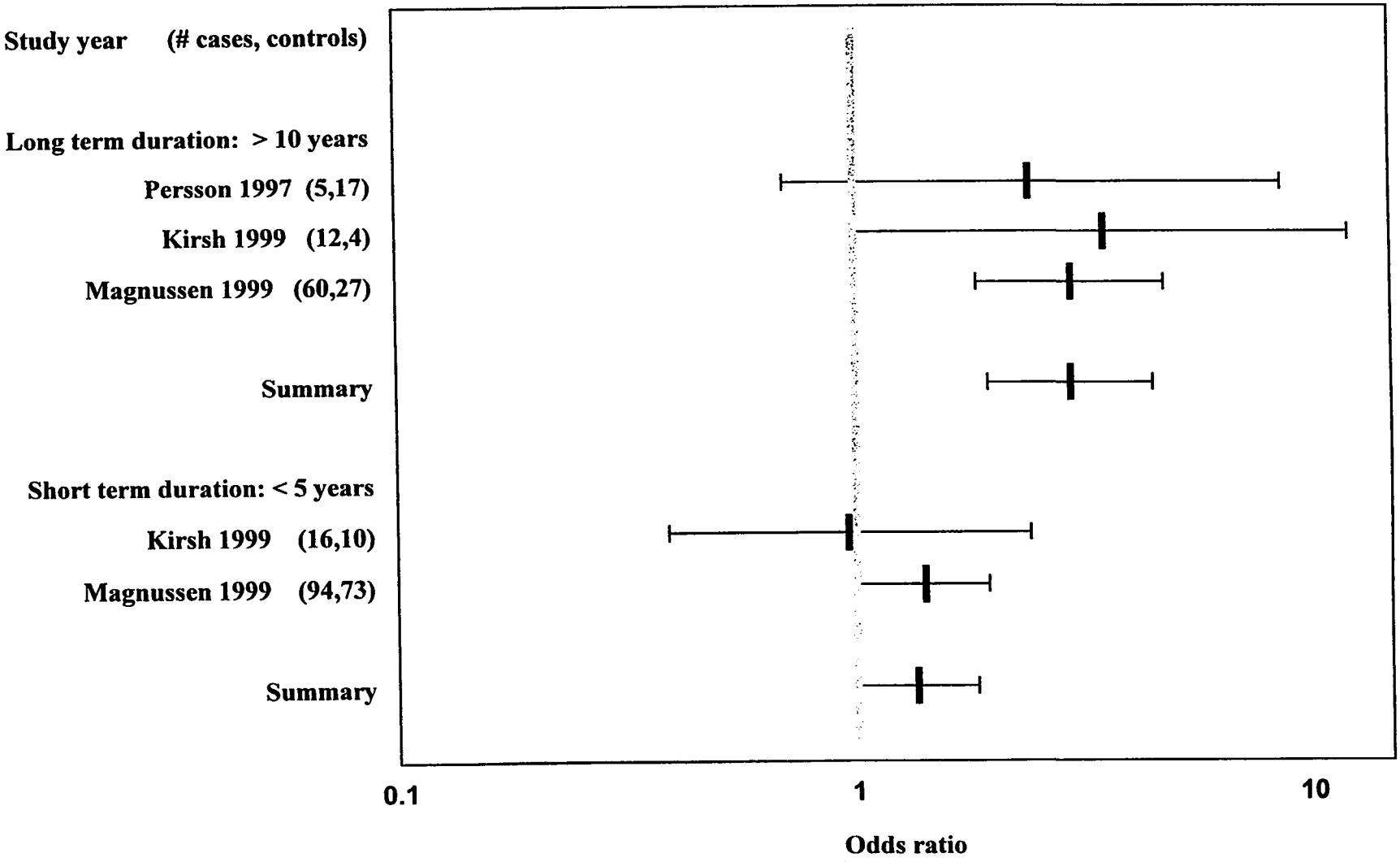
different lengths of exposure to EPRT. For instance, both Stanford et al[115] and Newcomb et al[111] provided estimates of the breast cancer risk for five different durations and none of the exposure durations were for the same time period or length of time. Stanford's study reported on exposure durations of 1-3 months; 4 months-2.9 years; 3-4.9 years; 5-7.9 years and 8 years or more while Newcomb's study provided details on exposure durations less than 2 years; 2-4 years; 5 to 9 years; 10-14 years and 15 years or more. Given the variation in the exposure durations that were reported, it would not be possible to compare these results in a meaningful way. It would not even be possible to simply combine some of the durations to get a more comparative estimate because the resulting OR would be unadjusted and thus would not control for the various risk factors (eg. age at menarche, BMI).

The more recent studies had more comparable time frames for exposure duration. Three studies reported on exposure duration for more than 10 years and two of these studies also reported on exposure duration for less than five years. A subgroup analysis comparing long and short term exposure to EPRT was performed using the fixed effects model. The ORs and summary estimates are plotted in Figure 4.8.

The results from the meta-analysis revealed an almost three fold increase in breast cancer risk for postmenopausal women exposed to EPRT for 10 years or longer (OR = 2.93; 95% CI 1.94-4.45; Table 4.8). The breast cancer risk for shorter term EPRT use was considerably less than the risk for longer term use and was not quite significant.

Chi-square tests of association and heterogeneity were performed for the two subgroups: long and short term EPRT exposure duration. The results of these statistical

Figure 4.8 Subgroup analysis of study duration: odds ratios and 95% confidence intervals for long vs. short term exposure to EPRT in case control studies.



tests of significance are displayed in Table 4.8. Two of the three studies showed a significant breast cancer risk associated with EPRT use. Overall, there was a significant association between breast cancer risk and long term use of EPRT. The chi-square test revealed that there was no significant heterogeneity between the three studies and thus it was reasonable to use the fixed effects model to arrive at the summary estimate.

The breast cancer risk associated with short term EPRT use was considerably less than the risk associated with long term use, and only marginally significant. Of the two studies entered into the analysis, only the larger one[74] showed a significant breast

Table 4.8. Results of the chi-square tests of association and homogeneity for subgroup analysis based on duration of EPRT exposure.

	Long term duration > 10 years	**Short term duration < 5 years**
Summary (95% CI)	2.93 (1.94-4.45)	1.34 (0.99-1.83)
# of studies	N = 3	N = 2
$X^2_{association}$	25.85 $p < 0.0001$	3.65 $p = 0.056$
$X^2_{homogeneity}$	0.17 $p = 0.919$	0.62 $p = 0.435$

cancer risk associated with EPRT use. There was no significant heterogeneity between the two studies.

It is also interesting to compare short and long term exposure duration within the same study. According to the Cochrane Reviewer's Handbook, Clarke and Oxman[82] report that "differences in subgroup response observed within a single study provide a stronger basis for drawing conclusions". This can be done with two of the studies: Kirsh[117] and Magnussen et al[132]. In the Kirsh study, the OR for long term EPRT use is

significant (OR = 4.03; 95% CI 1.11-14.64) while the one for short term use is not (OR = 0.95; 95% CI 0.38-2.4). In the Magnussen study, while both ORs are significant, the long term EPRT OR is much larger (OR = 2.95; 95% CI 1.84-4.72) than the short term one (OR = 1.40; 95% CI 1.01-1.94) and there is only a slight overlap in the estimates. The results of these two case control studies indicate that there is greater breast cancer risk associated with long term exposure (greater than 10 years) to EPRT and there is less of a risk associated with short term exposure.

The relationship between breast cancer risk and duration of exposure to EPRT was also examined in the cohort studies. None of studies had comparable exposures and thus a comparative analysis could not be performed between studies. However, within study comparisons were possible for three of the four cohort studies. Colditz et al[71] stated that there were too few women in their study taking EPRT to permit a detailed analysis of the risk according to duration of use.

The breast cancer risk for three different exposure durations (2-4 years; 5-10 years and greater than 10 years) was assessed in the Laurtizen and Meier[36] study. While none of the estimates were significant, they reported that the RR increased as the exposure duration increased (from 0.88 to 0.98 to 1.10).

Persson et al[38] compared the breast cancer risk for two different durations of EPRT exposures: 1-6 years and greater than 6 years. There was a significant breast cancer risk reported for women that were exposed for the longer duration (RR = 1.7; 95% CI 1.1-7.6) but no significant risk for the women exposed for the relatively shorter duration (RR = 1.4; 0.9-2.3).

Schairer et al[39] also compared the breast cancer risk for two different durations of EPRT exposures: less than 4 years and 4 years or more, and came up with similar findings to the Persson study (Figure 4.9). Schairer et al reported a significant risk associated with long term exposure to EPRT (RR = 1.9; 95% CI 1.1-3.3; n = 20 cases), but no significant risk associated with short term exposure to EPRT (RR = 1.1; 95% CI 0.7-1.8; n = 18 cases).

While the evidence from these two cohort studies could not be combined because the exposure durations differed, their results are consistent with the findings of the case control studies; namely that there is an increased breast cancer risk associated with long term exposure to EPRT.

4.4.2.3 Current use of estrogen progestin replacement therapy

Most of the studies in the meta-analysis compared ever use of EPRT with never use of HRT. Some of the more recent studies examined the influence of current EPRT use on breast cancer to determine if recency of use affected the risk of breast cancer. Six of the studies included in the meta-analysis (three case control studies, three cohort studies) provided estimates for the breast cancer risk among postmenopausal women that were classified as current users of EPRT. A current user was defined in all of these studies as someone that had been on EPRT within the past 12 months.

The results from the three case control studies[116;117;132] revealed that there was a significant breast cancer risk among the current users of EPRT (Table 4.9). The overall summary estimate indicates that there is a significant breast cancer risk associated with current use (Figure 4.10). However, this risk was not significantly different from the risk

Figure 4.9 Subgroup analysis of study duration: relative risks and 95% confidence intervals for long vs. short term exposure to EPRT in cohort studies.

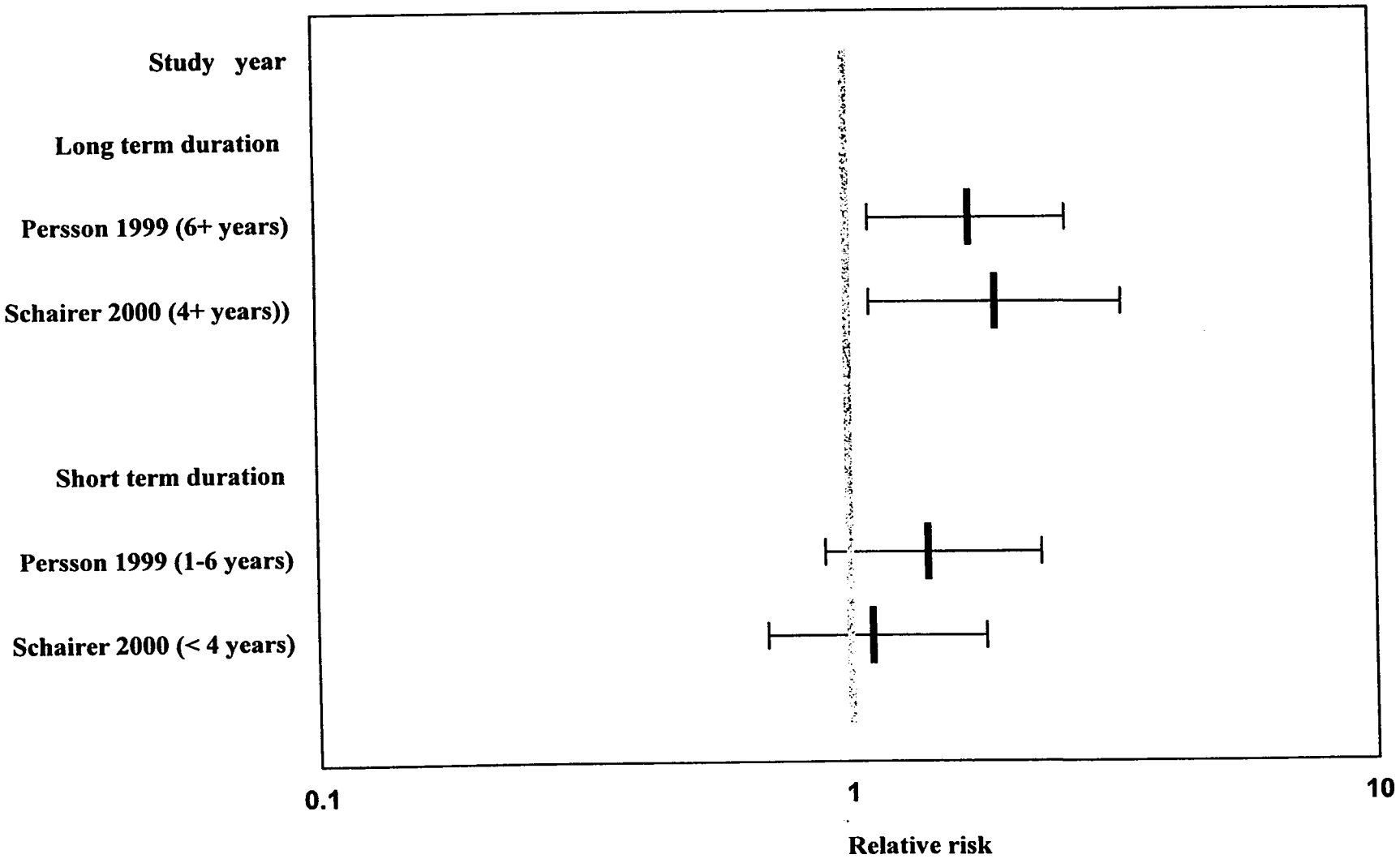
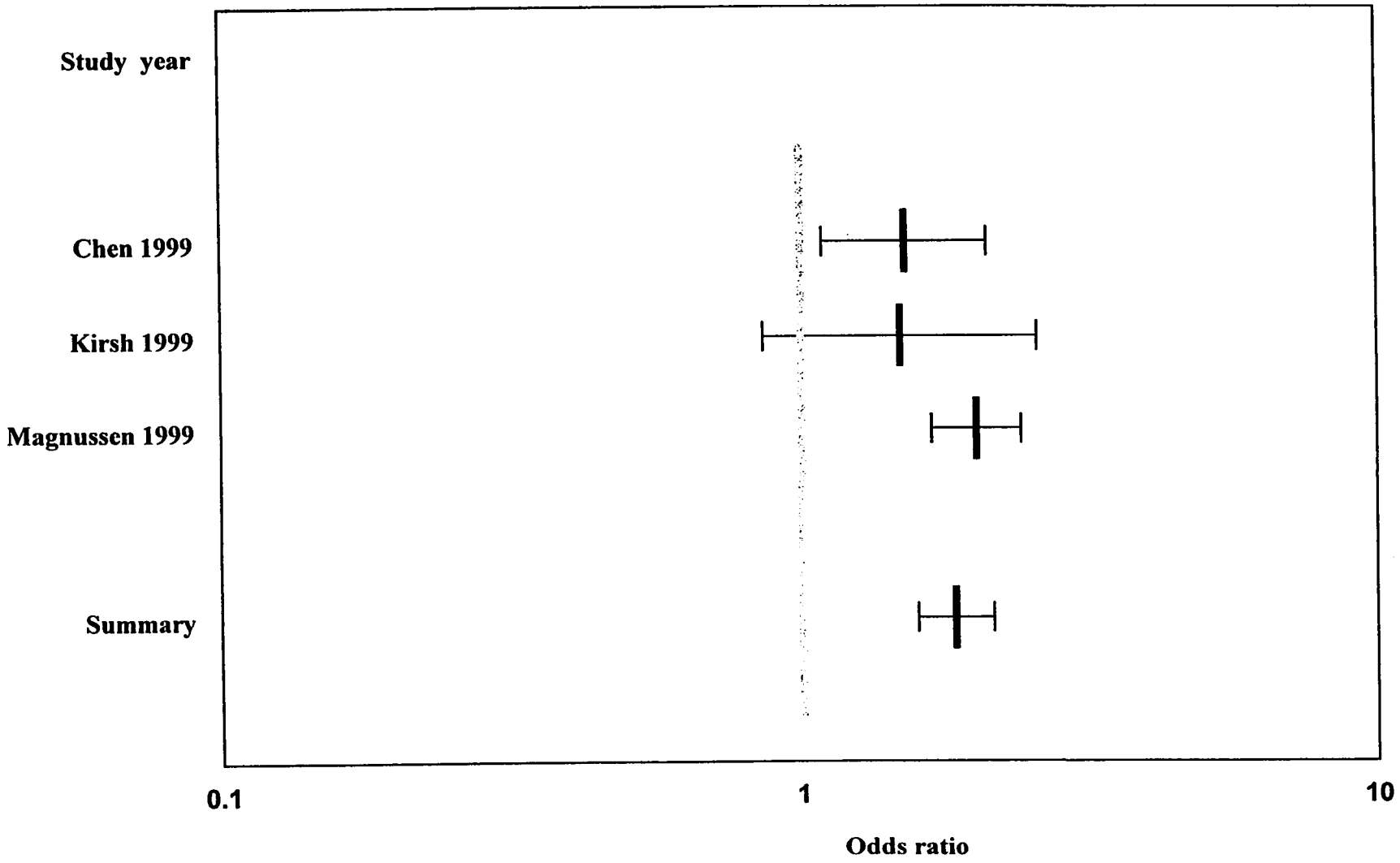


Figure 4.10 Subgroup analysis of current EPRT use: odds ratios and 95% confidence intervals of current EPRT use vs. never use of HRT in case control studies.



that was found among ever users in these same three case control studies (OR = 1.57; 95% CI 1.35-1.82).

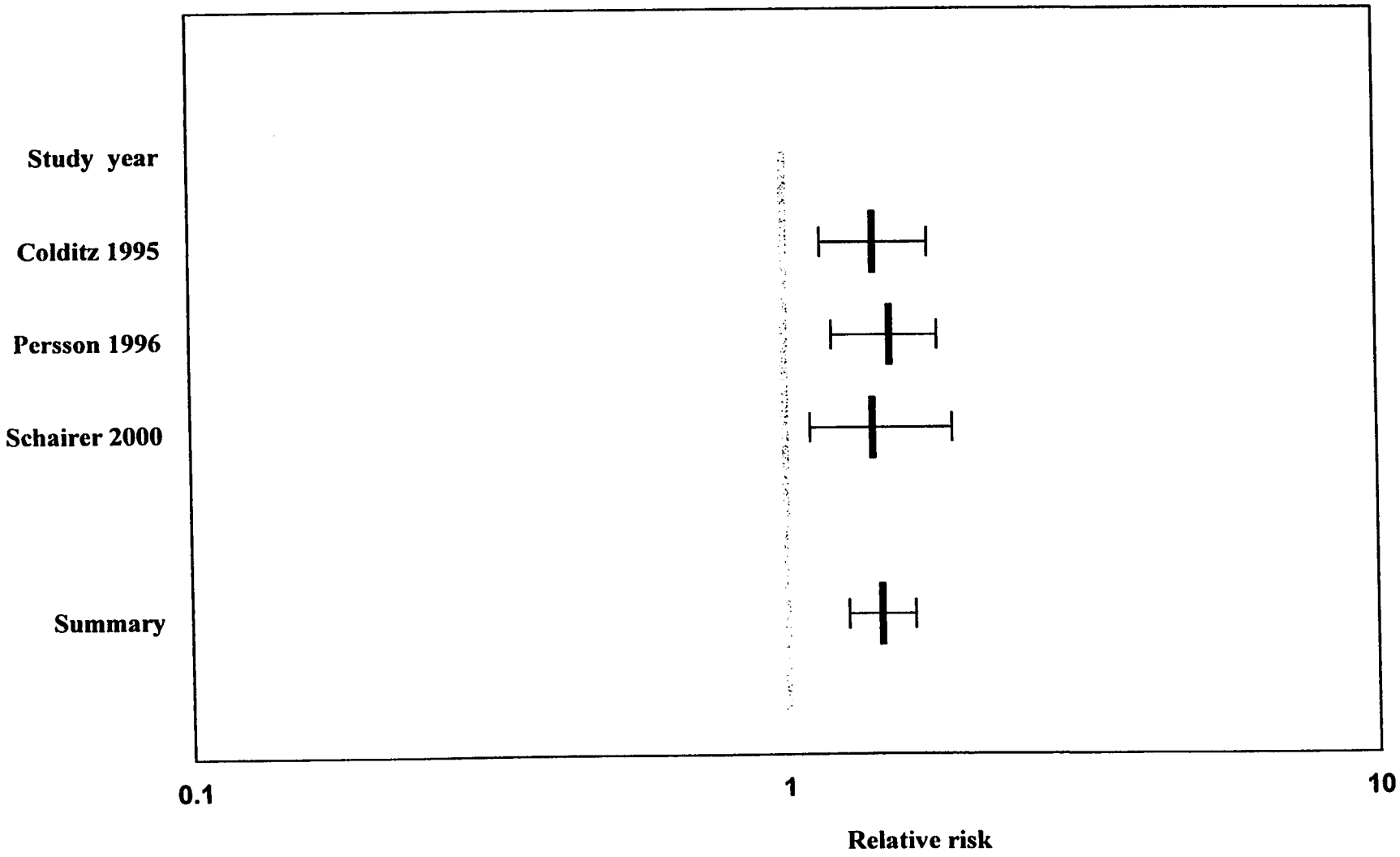
A similar result was noted for the cohort studies (Table 4.9, Figure 4.11). It is interesting to note the similarity in the risk estimates from these three studies[39;71;127]. The summary estimate indicates that there was a significant risk for current users of EPRT. However, this estimate was not significantly different from the one calculated for ever users in the same three cohort studies (Table 4.6, OR = 1.41; 95% CI 1.25-1.60).

The chi-square tests of association and homogeneity for current use of EPRT vs. never use of HRT are shown in Table 4.9. The results indicate that there was a significant association for current users of EPRT in both the case control and cohort studies and the heterogeneity between studies was not significant. It is difficult to assess whether or not current users are at a higher risk than ever users because not all of the studies reported data for current users. The available data suggest that there is a significant association between current EPRT use and breast cancer, but this relationship does not appear to differ from that observed between ever use of EPRT and breast cancer.

Table 4.9 Results of the chi-square tests of association and homogeneity for subgroup analysis based on current use of EPRT vs. never use of HRT

	Case control studies	**Cohort studies**
Summary (95% CI)	1.83 (1.58-2.13)	1.44 (1.27-1.64)
# of studies	N = 3	N = 3
$X^2_{association}$	63.43 $p < 0.0001$	31.53 $p < 0.0001$
$X^2_{homogeneity}$	2.81 $p = 0.25$	0.24 $p = 0.89$

Figure 4.11 Subgroup analysis of current EPRT use: relative risks and 95% confidence intervals of current EPRT use vs. never use of HRT in cohort studies.



4.4.2.4 Study location

Among the 11 case control studies, there were four from Europe and seven from North America. The ORs from each group were plotted in Figure 4.12 and their summary estimates were determined using the fixed effects model. The results from this subgroup analysis (Table 4.10) indicate that while there was a significant breast cancer risk associated with EPRT use in both locations, the risk was higher among European studies. There was overlap between the two estimates such that the lower 95% CI of the European studies was within the same range of values as the upper 95% CI of the North American studies.

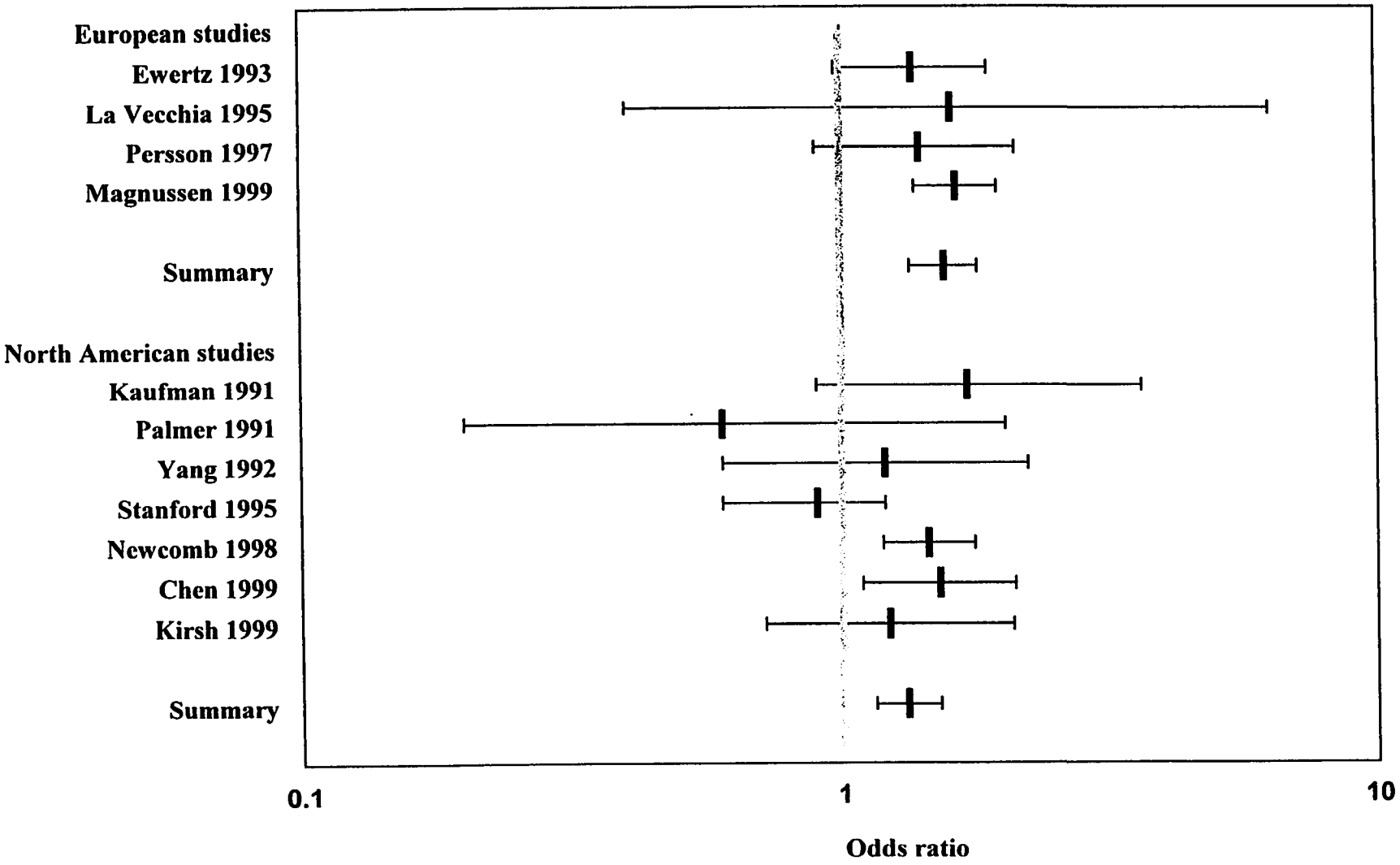
Table 4.10. Results of the chi-square tests of association and homogeneity for subgroup analysis based on study location.

	European studies	**North American studies**
Summary (95% CI)	1.55 (1.34-1.78)	1.32 (1.15-1.51)
# of studies	N = 4	N = 7
$X^2_{association}$	35.49 $p < 0.0001$	15.58 $p < 0.0001$
$X^2_{homogeneity}$	1.17 $p = 0.76$	8.62 $p = 0.2$

4.4.2.5 Testosterone vs. progesterone derived estrogen progestin replacement therapy

One of the reasons that a higher risk among European studies was observed may be due to differences in the progestins used in the EPRT formulation. There was not much variation in the type and dosage of estrogen used. Of the five studies reporting

Figure 4.12 Subgroup analysis of study location: odds ratios and 95% confidence intervals for European vs. North American case control studies.



specific formulation details, all reported that the estrogen in the EPRT consisted of a daily dose of 0.625 mg of conjugated estrogens. However, there was variation in the progestins used in the EPRT formulations.

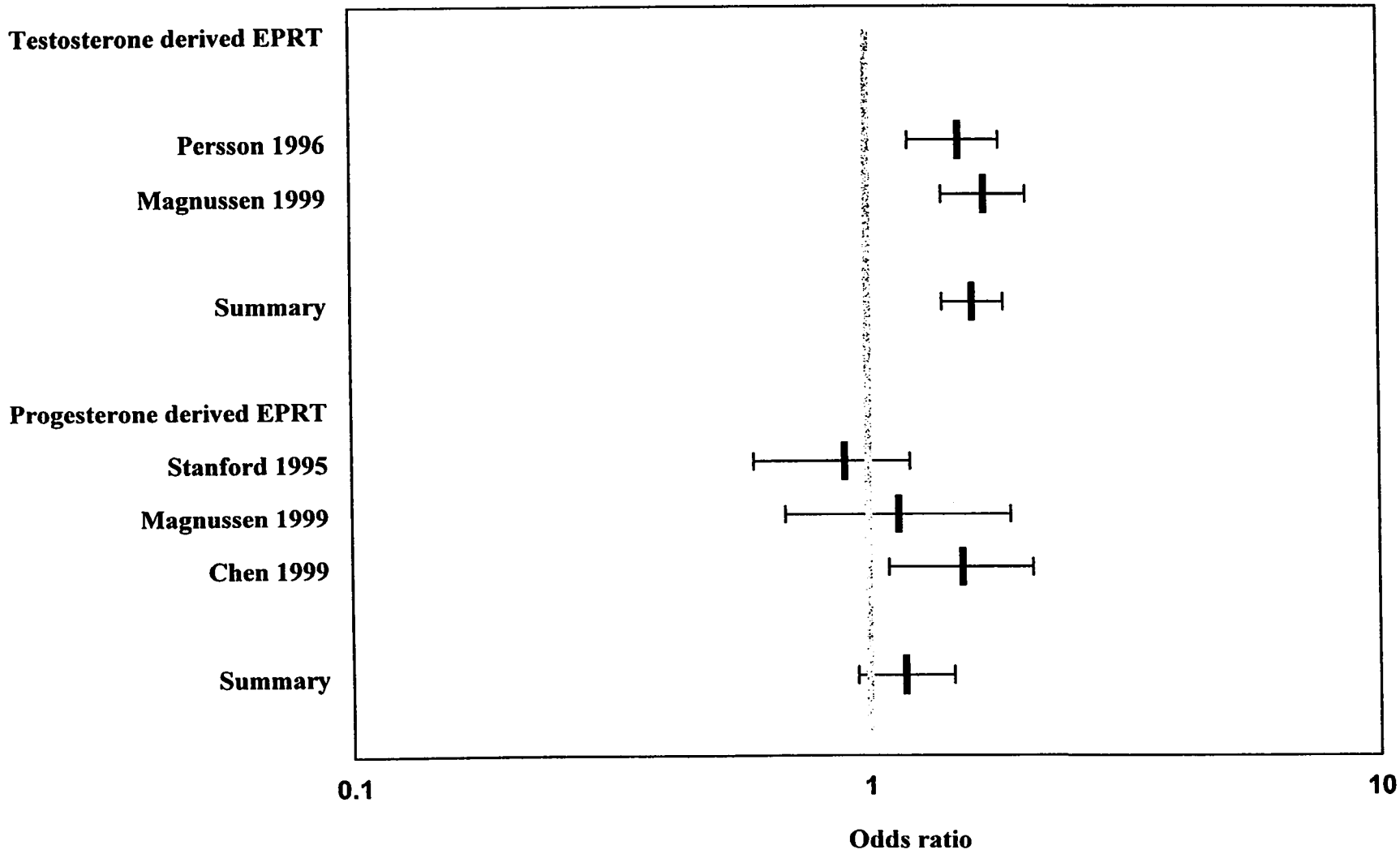
North American prescriptions tended to be mainly progesterone derived while there was more variation in EPRT formulation within the European studies. Only two of the four European studies provided specific details on EPRT formulation. For instance, Magnussen et al[74] reported that 79% of the postmenopausal women taking EPRT (n = 703) were on a testosterone derived formulation. Persson et al[38] reported that 45% of the EPRT formulations were testosterone derived, while the remaining 55% were progesterone derived. The types of testosterone tended to be 19-nortestosterone derivatives such as norethisterone, norethisterone acetate, levonorgestrel or lynestrenol.

None of the North American studies reported use of testosterone derived EPRT. Of the five North American studies providing details on EPRT formulations, all reported that the EPRT formulations were progesterone derived and identified the 17-hydroxyprogesterone derivative as medroxyprogesterone acetate.

To determine the affect of progestin-type on breast cancer risk, a separate analysis was completed comparing the two different progestin formulations. One European study[74] provided data for women on both types of EPRT. The OR values were plotted in Figure 4.13 and the summary estimates for the two groups were determined using a fixed effects model.

There was a significant breast cancer risk associated with exposure to testosterone derived progestins, but no significant risk associated with exposure to progesterone derived progestins. There was only a very slight overlap in the two summary estimates,

Figure 4.13 Subgroup analysis of EPRT formulation: odds ratios and 95% confidence intervals for testosterone vs. progesterone derived EPRT case control studies



suggesting that use of testosterone derived formulations clearly poses a higher risk of breast cancer. When comparing the results within the same study[74], the same pattern emerges. Magnusson et al. found that women on testosterone derived EPRT had a significant breast cancer risk (OR = 2.31; 95% CI 1.37-3.88) while women on the progesterone derived formulation had a lower and nonsignificant risk (OR = 1.14; 95% CI 0.69-1.88).

The chi-square tests of association and homogeneity for progestin type are shown in Table 4.11. A significant association for exposure to testosterone derived progestins was observed and the heterogeneity between studies was not significant. However, there was no significant association observed for exposure to progesterone derived progestins and no significant heterogeneity between studies.

Table 4.11. Results of the chi-square tests of association and homogeneity for subgroup analysis based on type of progestin used in the EPRT formulation

	Testosterone derived EPRT	**PROGESTERONE DERIVED EPRT**
Summary (95% CI)	1.59 (1.39-1.83)	1.17 (0.95-1.45)
# of studies	N = 2	N = 3
$X^2_{association}$	43.56 $p < 0.0001$	2.17 $p = 0.14$
$X^2_{homogeneity}$	0.64 $p = 0.42$	4.71 $p = 0.10$

The fact there was no significant heterogeneity between studies within the two progestin formulations suggests that some of the variation observed between the 11 case control studies may be explained by difference in the formulations. The fact that testosterone derived formulations tend to have a higher breast cancer risk may explain why the European studies in the previous subgroup analysis had a higher risk than the

North American studies. Unfortunately, this hypothesis cannot be tested directly because most of the studies did not describe the exact formulation of the EPRT used.

4.4.3 Sensitivity analysis

4.4.3.1. Case control studies

In order to determine the effect of including or excluding studies on the results of the meta-analysis, two sets of sensitivity analyses were performed. The summary estimates stemming from the sensitivity analyses for case control studies are plotted in Figure 4.14 and for cohort studies in Figure 4.15. The results of the analyses to test the significance of the association between breast cancer risk and EPRT exposure and to determine if there is significant heterogeneity between studies is reported in Tables 4.12 and 4.13.

The first series in the case control sensitivity analyses involved the inclusion of two studies that were originally excluded from the analysis: Hunt et al[122] and Ross et al [40]. An unadjusted OR and 95% CI were calculated from the data provided in Hunt (OR = 1.55; 95% CI 0.79-3.1). An unadjusted 95% CI was calculated from the data provided in Ross et al (OR = 1.18; 95% CI 0.99-1.40). Both of these were entered into the calculation of the summary estimates along with the original 11 studies that met all of the inclusion criteria. The resulting summary estimate does not differ from the estimate excluding these two studies. The association between breast cancer risk and EPRT exposure remains significant and there is no significant heterogeneity between these 13 studies.

Figure 4.14 Sensitivity analysis: summary estimates resulting from changes in studies included/excluded from analysis of case control studies

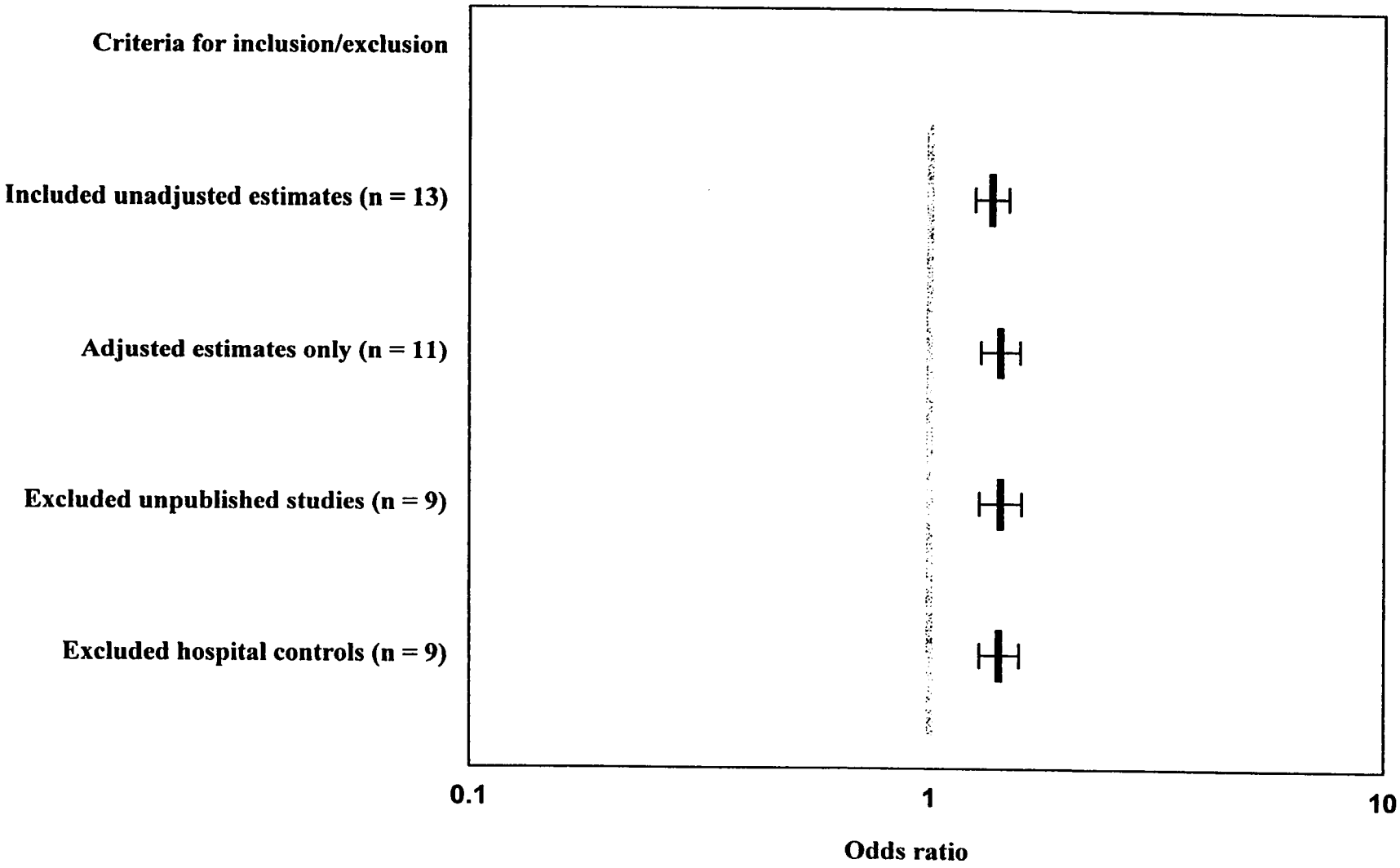
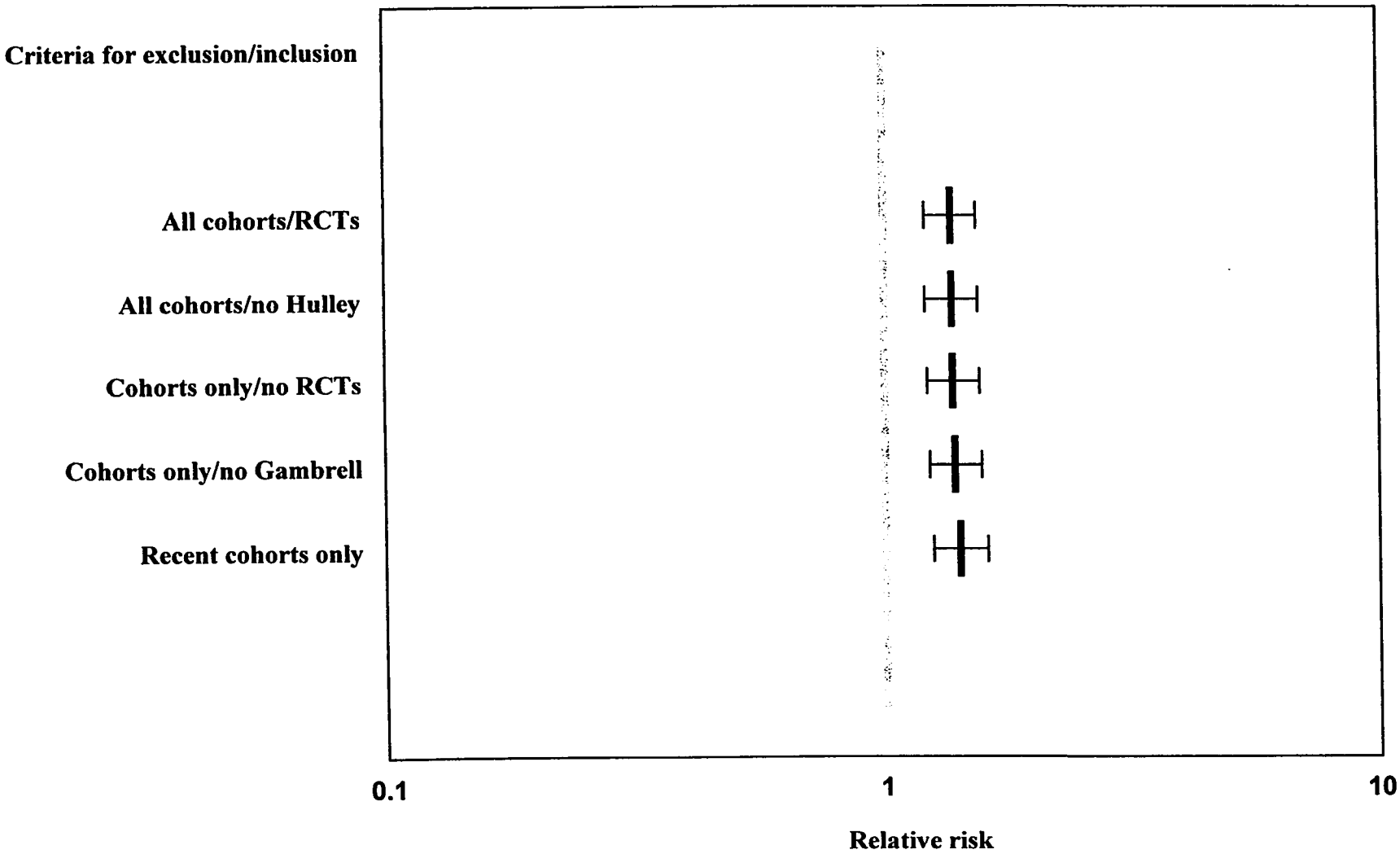


Figure 4.15 Sensitivity analysis: summary estimates resulting from changes in studies included/excluded from analysis of cohort studies



The next series involved the exclusion of studies that were not peer-reviewed. Two of the studies[116;117] were retrieved from Dissertation Abstracts. When they were excluded from the analysis, there was little change in the summary estimate. Furthermore, the chi-square tests revealed that the association remained significant and there was no heterogeneity between the remaining studies.

The final series involved excluding studies that recruited their controls from the hospital rather than the community. This can often introduce a bias into the study and thus by removing these two studies[109;120], the impact could be assessed. There was only a slight change in the summary estimate as a result of excluding these two studies. The chi-square for association remained significant and as with the other sensitivity analyses, there was no significant heterogeneity between studies (Table 4.12).

Table 4.12. Sensitivity analysis for case control studies: summary estimates of breast cancer risk associated with ever use of EPRT vs. never use of HRT among postmenopausal women

Studies included in the analysis	Number Of Studies	Summary OR (95% CI)	$\chi^2_{association}$ (p value)	$\chi^2_{homogeneity}$ (p value)
All case control studies[1]	N = 13	1.36 (1.25-1.48)	50.07 ($p < 0.0001$)	15.84 ($p = 0.20$)
Case control studies with adjusted risk estimates[2]	N = 11	1.42 (1.29-1.57)	48.30 ($p < 0.0001$)	12.32 ($p = 0.27$)
All peer reviewed studies with adjusted risk estimates[3]	N = 9	1.42 (1.28-1.58)	41.96 ($p < 0.0001$)	11.86 ($p = 0.16$)
All peer reviewed studies with adjusted risk estimates and community controls[4]	N = 7	1.41 (1.28-1.56)	45.90 ($p < 0.0001$)	12.32 ($p = 0.27$)

[1]including Hunt 1987; Ross et al 2000
[2]excluding Hunt 1987; Ross et al 2000
[3]excluding Chen 1999 and Kirsh 1999
[4]excluding Kaufman et al 1991; La Vecchia et al 1995

4.4.3.2 Cohort studies

The first series in the cohort sensitivity analyses (Table 4.13) involved the inclusion of the two RCTs that were originally excluded from the analysis: Hulley et al[19] and Nachtigall et al.[133] While the breast cancer risk estimate was provided in the Hulley study (RR = 1.30; 95% CI 0.77 – 2.19), one had to be calculated from the data provided in the Nachtigall study. Since there were no breast cancer cases among women on EPRT in this study, 0.5 was added to each of the cells in order to calculate the risk (RR = 0.11; 95% CI 0.01 – 2.00). Both of these estimates were entered into the calculation of the summary estimate along with the five cohort studies.[36;37;39;71;127] The resulting summary estimate does not differ from the summary estimate excluding these two studies or from the summary estimate that included only those studies that met all of the inclusion/exclusion criteria. The association between breast cancer risk and EPRT exposure remains significant and there is no significant heterogeneity between these seven studies.

The next step in the sensitivity analysis involved excluding the Hulley study (it included women that had prior exposure to ERT) and including only one RCT, the Nachtigall study. Once again, this made no difference in the summary estimate or in the statistical analysis, further demonstrating the robustness of the results.

The removal of the Gambrell study[37] was the next exclusion in the sensitivity analysis (Table 4.13). It was initially excluded because no adjusted risk estimate was provided. As was the case with the earlier steps in the sensitivity analysis, this had no bearing on the final results.

The final step in the sensitivity analysis involved the exclusion of the Laurtizen and Meier study[36], the only study that was not peer-reviewed. The results of this study were published as a chapter in a book. This exclusion had little impact on the summary estimate and the chi-square tests revealed that the association remained significant and there was no heterogeneity between the remaining studies.

Table 4.13. Sensitivity analysis for RCTs and cohort studies: summary estimates of breast cancer risk associated with ever use of EPRT vs. never use of HRT among postmenopausal women

Studies included in the analysis	Number of Studies	Summary OR (95% CI)	$\chi^2_{association}$ (p value)	$\chi^2_{homogeneity}$ (p value)
All cohort and RCT studies[1]	N = 7	1.36 (1.21-1.53)	28.19 (p < 0.0001)	9.34 (p = 0.16)
All cohort and RCTs satisfying inclusion criteria[2]	N = 6	1.37 (1.21-1.54)	26.26 (p < 0.0001)	9.30 (p = 0.27)
All cohort studies[3]	N = 5	1.37 (1.22-1.55)	27.09 (p < 0.0001)	5.72 (p = 0.22)
All cohort studies with adjusted risk estimates[4]	N = 4	1.38 (1.23-1.56)	27.95 (p < 0.0001)	2.81 (p = 0.42)
Peer-reviewed cohort studies with adjusted risk estimates[5]	N = 3	1.41 (1.25-1.60)	29.92 (p < 0.0001)	0.82 (p = 0.67

[1]including Hulley et al[19]; Nachtigall et al[130]
[2]excluding Hulley et al(106)
[3]excluding Nachtigall et al[130]
[4]excluding Gambrell[37]
[5]excluding Lauritzen and Meier[36]

4.4.4 Cumulative meta-analyses

Another type of sensitivity analysis, cumulative meta-analysis, was carried out in order to determine if the relationship between breast cancer risk and EPRT exposure remained robust under different conditions. In this study, two different cumulative meta-analyses were performed. The first type involved entering the studies chronologically to see if the results changed over time. The second type involved entering the studies in

order of study quality to see if the results remained consistent with a decline in study quality.

The summary estimates that were sequentially computed by adding one new study at a time in a predetermined order are plotted in Figures 4.16-4.19. In these figures, each horizontal line represents the summary estimate of the results as each study is added (a cumulative summary), rather than the results of a single study. The study names along the vertical axis represent the order in which the studies were added. For instance, in Figure 4.16, Palmer[118] was the first study entered. The next entry was Kaufman[120] and the summary estimate represents the combination of the ORs from these two studies using the fixed effects model. The third line in the figure, "P-K/Yang 1992" indicates that the results from Yang[134] has been combined with the results from Palmer and Kaufman to derive a third summary estimate. As one moves down the graph, more studies are sequentially entered.

The results of the chronological cumulative meta-analysis for case control studies (Figure 4.16) indicate that, over time, the breast cancer risk associated with EPRT exposure increases and changes from being non-significant to significant. A similar trend was also observed when the comparative analysis was completed for cohort studies (Figure 4.17). The breast cancer risk associated with EPRT exposure increased and became significant as more recent studies were incorporated into the analysis.

Interesting trends were also observed for the cumulative meta-analysis for study quality. Once again, the results observed for case controls studies (Figure 4.18) and cohort studies (Figure 4.19) showed comparable trends. The studies in these two meta-analyses were incorporated in order of study quality, from the highest to the lowest NOS

Figure 4.16 Cumulative meta analysis: chronological assessment of case control studies

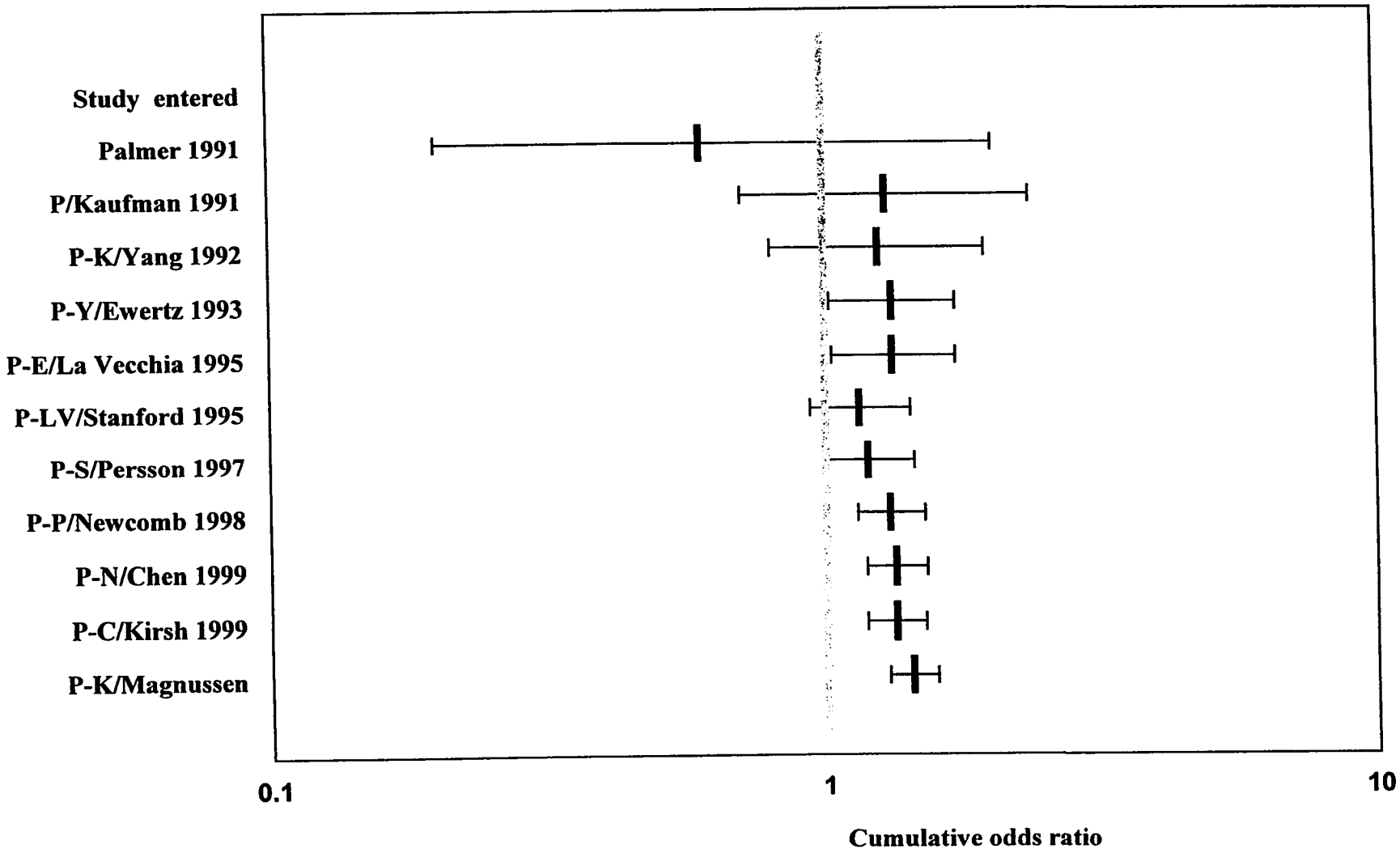


Figure 4.17 Cumulative meta analysis: chronological assessment of cohort studies

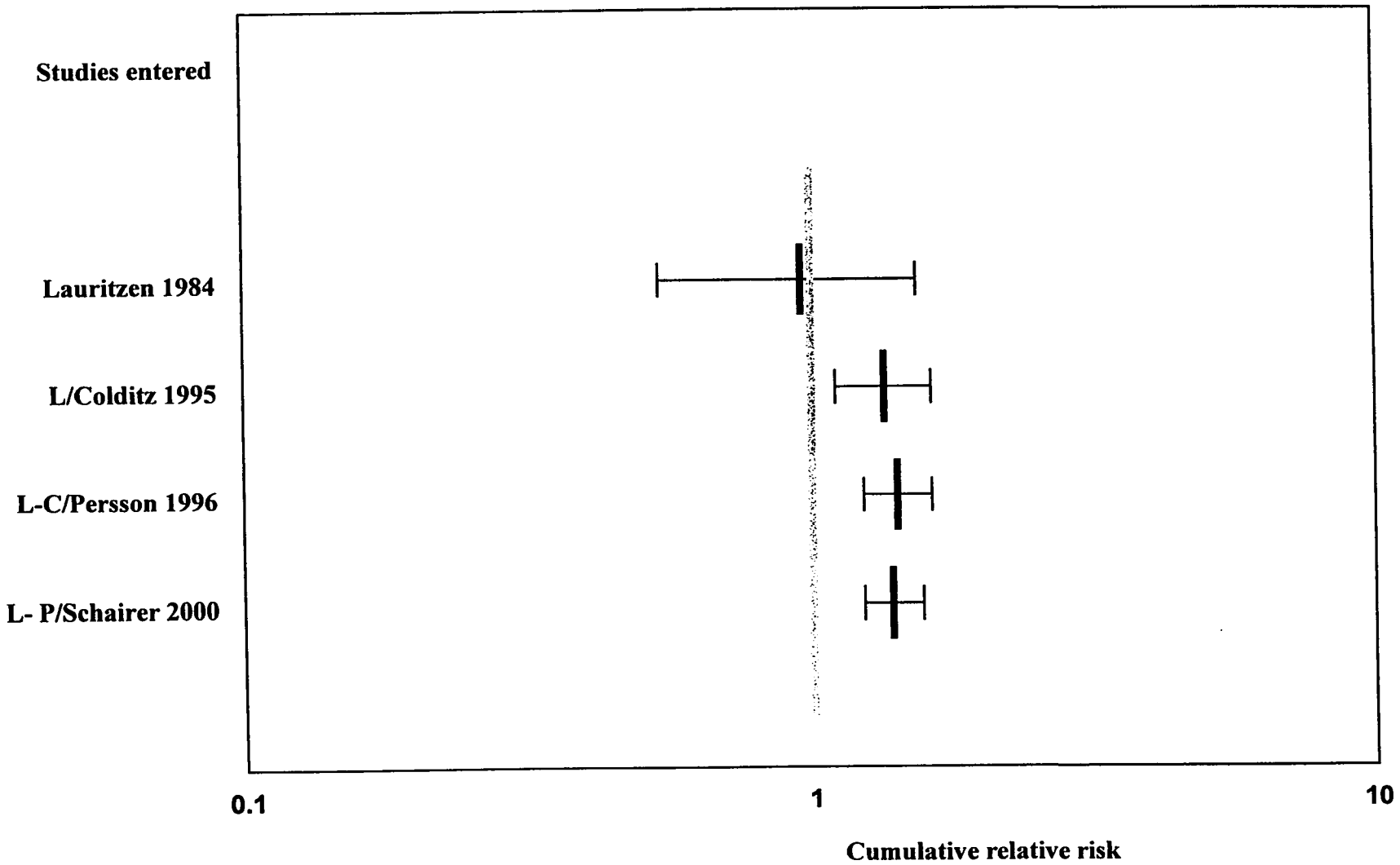


Figure 4.18 Cumulative meta analysis: quality assessment of case control studies

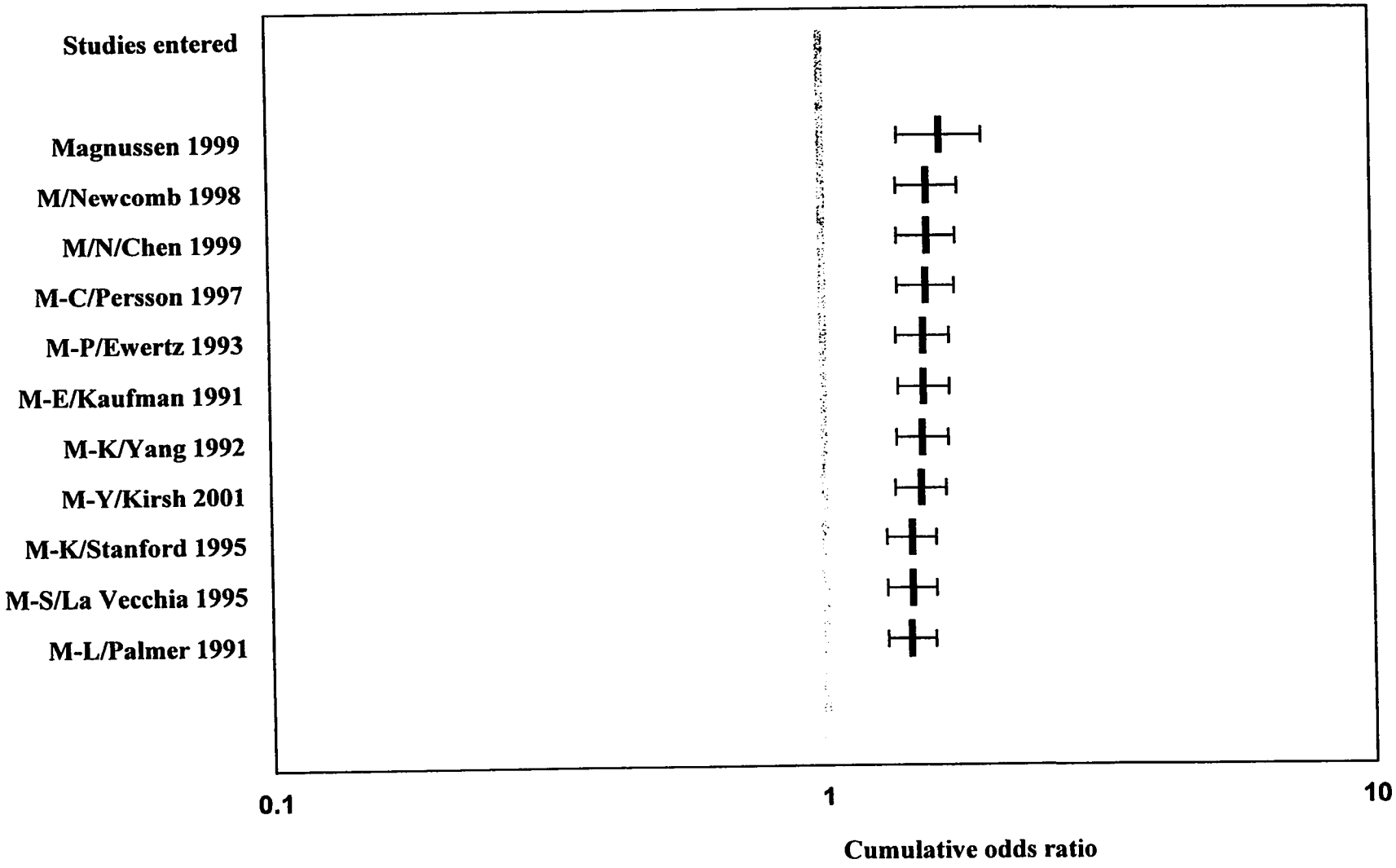
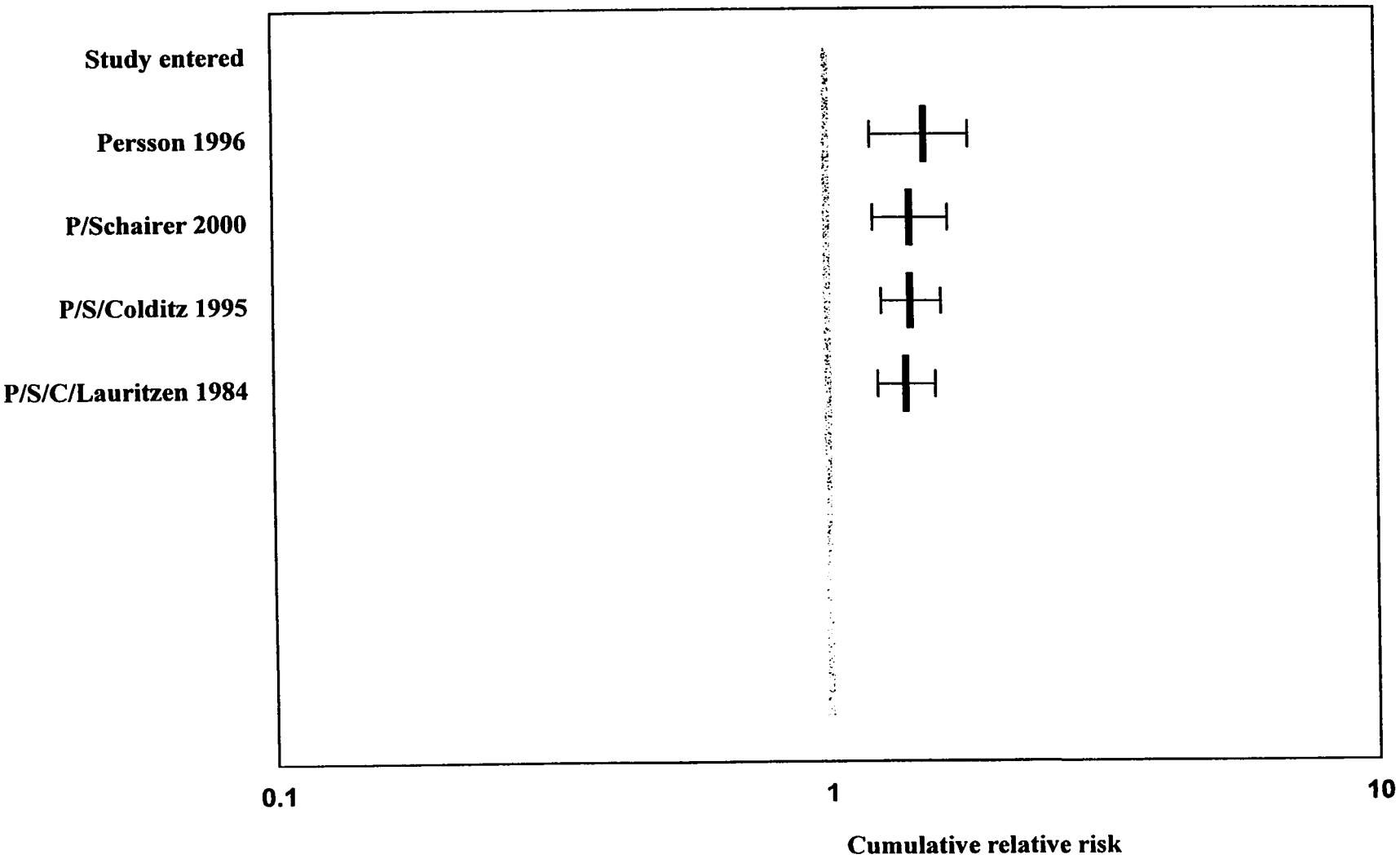


Figure 4.19 Cumulative meta-analysis: quality assessment of cohort studies



score. It is important to note that the range of study quality was not extreme (for case control studies, the NOS scores ranged from eight to six and for cohorts, scores ranged from eight to five) and thus extreme changes would not be expected.

The results from both study quality meta-analyses indicate that as the study quality declines, the estimate of the breast cancer risk associated with EPRT use also declines, albeit slightly. It is also important to note that even though the risk was observed to decrease, it remained significant.

4.4.5 Publication bias

The purpose of the funnel plot is to detect publication bias, defined as " the tendency on the parts of investigators, reviewers and editors to submit or accept manuscripts for publication based on the direction or strength of the study findings"[135]. However, for the purposes of this meta-analysis, it was difficult to decide what study findings should be used for the funnel plot construction.

All of the 13 published studies examined in this meta-analysis (nine case control studies; four cohort studies) provided information on the breast cancer risk associated with EPRT as well as unopposed ERT. In fact, for all but one study[115], the primary focus of the research was the relationship between ERT and breast cancer. The population of women on ERT was always larger than the population on EPRT and the abstracts always quoted the ERT data first. While it would be almost impossible at this point to determine which data influenced the decision for publication, it poses an interesting dilemma when attempting to construct funnel plots.

In an attempt to solve this problem, two funnel plots were constructed—one using the EPRT data set (Figure 4.20) and the other using the ERT data set (Figure 4.21). Originally, separate plots were constructed for case control and cohort studies. However, it was difficult to assess the shape of the cohort plots with only four data points. Therefore the cohort and case control data were combined into a single funnel plot.

According to Clark and Oxman[82], in the absence of bias, the plot of effect vs. treatment size should resemble a "symmetrical inverted funnel". While the shapes of the two resulting funnel plots were similar, the ERT data was more centered around the y axis. Both plots had a single study that was characterized by a large study size and intermediate effect size, representing the top of the funnel. The majority of the studies in both plots fell within the area characterized by a smaller study size and more extreme effect size. This formed the base of the funnel.

Figure 4.20 Funnel plot to assess publication bias in case control and cohort studies using EPRT data
Standard error vs. log odds ratio/relative risk

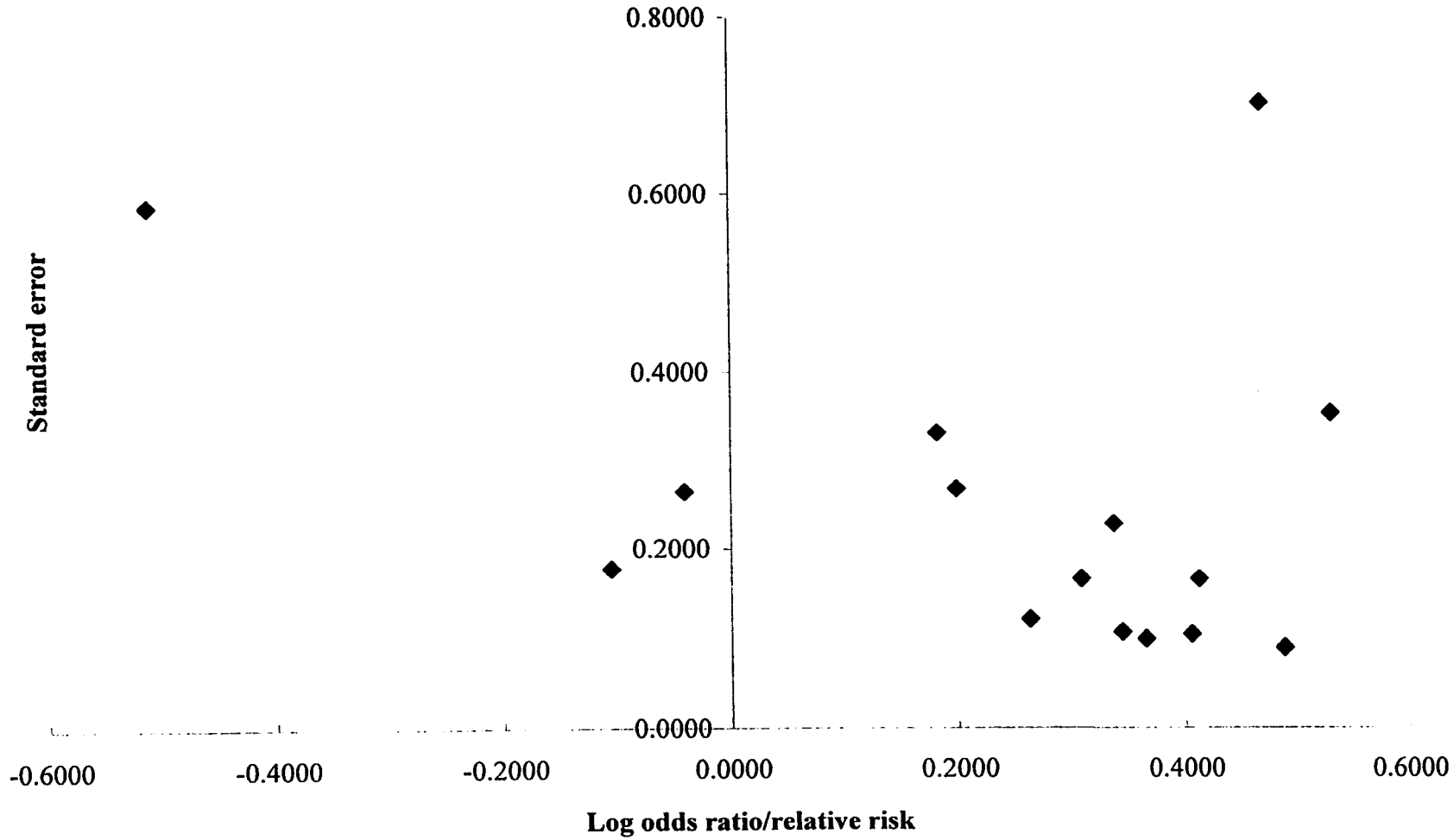
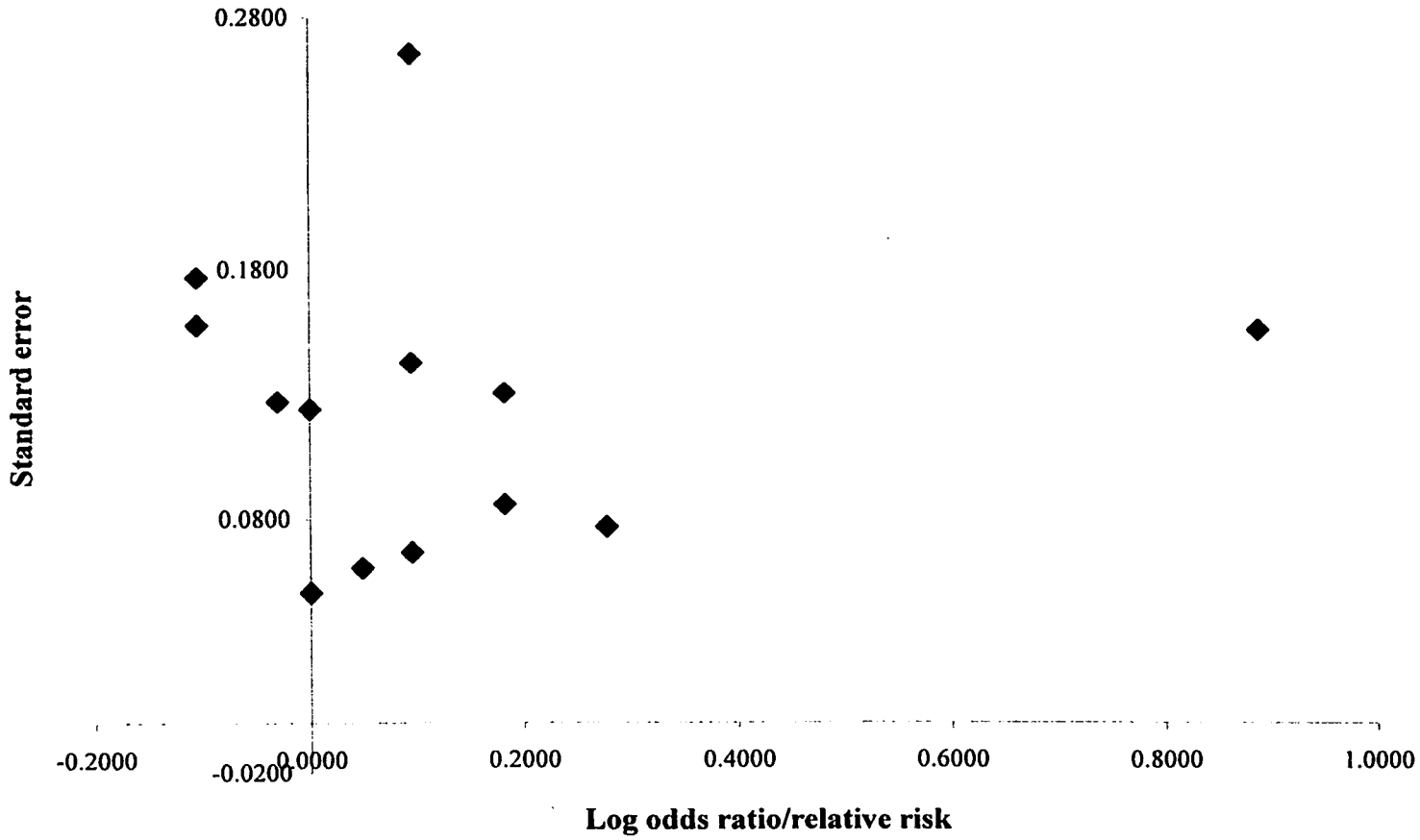


Figure 4.21 Funnel plot to assess publication bias in case control and cohort studies using ERT data standard error vs. log odds ratio/relative risk



5. DISCUSSION

The results of this study indicate that postmenopausal women exposed to EPRT have a thirty to forty percent higher risk of breast cancer compared to unexposed postmenopausal women. In this chapter, the search strategy employed to identify the pertinent literature will be assessed. The characteristics of studies used in the analysis will be examined and biases inherent in these studies will be discussed. The NOS tool, used to assess the quality of these studies, will also be reviewed. The results of the meta-analysis will be discussed in light of current EPRT and ERT literature and the significance of the findings will be examined. Future research directions will also be identified.

5.1 Search results

While there was at least a 29% overlap in the articles reported by the five main databases (Medline, Cancerlit, Current Contents, Biological Abstracts and Embase; see section 4.1.1), there were still sufficient differences in the articles to warrant the searching of all five. Although there was no monetary cost associated with seaching the first four databases, there was a considerable monetary cost associated with Embase.

Sampson et al[136] performed an analysis to determine whether excluding a search of Embase would introduce bias in the results of meta-analyses and therefore whether the added costs of an Embase search were warranted. They examined meta-analyses of RCTs that had been published since 1995. In their review of 98 meta-analyses, they found that the studies referenced exclusively through Embase tended to have smaller

estimates of intervention effectiveness. They argued that a review may be biased if Embase is not searched. However, there were very few Embase-unique trials revealed by their search and from this, they concluded that the risk of bias resulting from an exclusion of Embase is minimal, providing the rest of the search is comprehensive. Similar to the findings of Sampson et al., the Embase search from this study revealed no unique case control or cohort studies that could be incorporated into the meta-analysis. However, it did provide 17 unique references to other articles that were useful in preparing this thesis. Whether or not this made a difference in the quality of this systematic review is difficult to assess. However, it did provide assurance that a thorough search of all pertinent databases was completed.

There were 1631 articles from the search results that did not meet the inclusion/exclusion criteria. There are a number of modifications that could be introduced into future search strategies to reduce the number of unrelated articles, thus reducing screening time. For instance, many of the articles dealt with SERMs (selected estrogen receptor modulators) and were not relevant to the study question. To exclude these citations from the search results, simple limits could be introduced into the search strategy that would identify all of the SERM-related citations and then remove them from the search results. Other limits could be placed in the strategy that would restrict the results to only human studies, thus excluding the animal studies.

In order to reduce retrieval bias, Felson[98] recommends employing a variety of sources, in addition to electronic databases. In this study, the references from all retrieved articles were reviewed. While this did not yield any new studies, it did provide

additional background information and identification of laboratory and clinical studies that provided support to the epidemiological findings.

Handsearching of the current literature was also a useful exercise for revealing recent studies that examined health benefits and risks associated with HRT use. While this exercise did not disclose any new studies examining the association between EPRT and breast cancer, it did provide a recent cohort study examining the relationship between HRT and breast cancer[106]. It also provided articles on the effects of EPRT on breast tissue[63;72] and recent reviews on the epidemiology of breast cancer[55] and treatments for menopausal symptoms[18].

Contact with authors of unpublished studies included in this meta-analysis was very useful. Both Chen[116] and Kirsh[117] provided additional details about their unpublished theses that permitted the inclusion of their results into this study. However, attempts at obtaining further study details from authors of published studies were not effective.

5.2 Selected studies

Traditionally, medical meta-analyses were performed using the results of randomized studies. The search strategy used for this thesis identified only two RCTs that examined the association between breast cancer and EPRT.[19;130] They were subsequently excluded from the meta-analysis because they either failed to control for important breast cancer risk factors or because a significant proportion of the placebo group had prior

exposure to ERT (section 4.2.3). The remaining studies that met the inclusion/exclusion criteria (section 3.2) were all nonrandomized studies.

Meta-analyses of nonrandomized studies are increasingly found in the literature. Petitti[80] found that over forty percent of the meta-analyses published in medical and speciality clinical journals from 1997-1998 were based on nonrandomized studies. Shapiro[137] identified several problems associated with such meta-analyses. In particular, he dismissed the meta-analyses of breast cancer and ERT use. He argued that, in spite of nonsignificant relative risks, authors still claimed that a positive association existed when the appropriate selection of studies were included in the meta-analysis. He recommended that meta-analyses of nonrandomized studies should be abandoned. However, Greenland[79]argued that a qualitative review is "no longer the most reliable method of summarizing research". He stressed that a meta-analysis of nonrandomized studies should contain both quantitative (calculation of summary estimates) and qualitative aspects (explanations of the observed heterogeneity).

There are several problems associated with nonrandomized studies that are not encountered with RCTs. These problems can potentially introduce bias into the results. In order to assess this bias, it is first necessary to briefly compare the differences between meta-analyses using RCTs and those using nonrandomized studies as their source of evidence.

According to the Cochrane Collaboration, there are four key components of a study that one must consider when doing a meta-analysis: the population of interest; the outcome; the study design; and, the interventions. When comparing RCTs and nonrandomized studies, there are obvious differences in study design. With meta-

analyses based solely on RCTs, the results from the studies could be readily combined, if the other three study components were comparable. In this investigation, there were two study designs used, case control and cohort studies. Because these two designs are quite different, the results from each type of study were examined separately. It is interesting to note that in this investigation, the findings of the case control and cohort studies were actually consistent, indicating that study design did not influence the conclusions. Given the similarity in these findings, this section will discuss the results of both designs concurrently, unless otherwise noted.

For an investigation of the association between breast cancer and EPRT exposure, there was very little difference between RCTs and nonrandomized studies in terms of two of the components: outcome (breast cancer) and population of interest (postmenopausal women). In this investigation, all of the case control studies and three of the four cohort studies provided explicit criteria for establishing the presence of breast cancer. The one cohort study where it was not clearly defined[36] involved the patients of the two authors. It is reasonable to assume that they would have an accurate assessment of their patients' health status and whether or not they had breast cancer.

The population of interest was postmenopausal women. In the 15 studies incorporated into this analysis, all of the women were classified as postmenopausal. There are often difficulties with this classification as it may include younger women that have had a premenopausal hysterectomy and thus are classified as having a surgical menopause. Since age is one of the most important risk factors for breast cancer, this could lead to an underestimate of the effect of menopausal age on breast cancer risk since surgically menopausal women would be younger[107]. This would be an important

consideration if this were a study of women on ERT, since women experiencing surgical menopause use ERT[39]. However EPRT is the standard therapy for women with an intact uterus; these women experience a natural menopause[138]. Therefore age at menopause would not have been underestimated in the EPRT studies incorporated in this investigation.

The only women taking EPRT who may have had an early menopause would be those that had breast cancer. Col et al[139]estimate that 70% of the women diagnosed with breast cancer during their reproductive years develop premature menopause as a result of adjuvant chemotherapy. To treat their menopausal symptoms, a small number of these women take EPRT[140]. This could introduce a bias similar to the one described above. However, this is not an issue for this investigation since all of the studies included in the meta-analysis excluded women with a prior history of breast cancer.

The component that presented the greatest potential for bias in nonrandomized studies involved the intervention. In this investigation, the intervention can be better expressed as the exposure of postmenopausal women to EPRT. In a typical RCT, postmenopausal women would be randomly allocated into two groups (treatment, placebo) to ensure that all other factors that may affect the risk of disease are controlled. For instance, the two groups would be similar in terms of their medical history, physical activity, smoking, alcohol consumption and other pertinent traits to ensure that the women in the two groups were at a similar risk for breast cancer. In addition, through random allocation, the treatment and placebo groups would be balanced in terms of unidentified risk factors, providing that the sample sizes are sufficiently large.

Nonrandomized studies would only be able to accommodate a few risk factors in their design and/or analysis through matching, restriction, stratification or statistical models.

In a properly designed RCT, the only difference between the placebo and treatment group would be their exposure to EPRT. Furthermore, the EPRT exposure among women within the treatment arm of the RCT would be identical. They would be exposed to the same EPRT formulation and dosage (eg. 0.625 mg of conjugated equine estrogen and 10 mg of medroxyprogesterone actetate) and for the same duration (eg. 10 years). Therefore, with a RCT, an accurate assessment of exposure could be readily determined, assuming that there is good participant adherence.

In nonrandomized studies, considerable variation in exposure could exist and one must rely on the accuracy of the participants' recall to gain insight into their exposure. Thus, it is more difficult to obtain an accurate assessment of exposure in nonrandomized studies. In the nonrandomized studies examined in this investigation, researchers compared exposures in terms of ever use of EPRT versus never use of HRT. Ever use implied use of EPRT for some unspecified length of time and at varying formulations and dosages. Never use indicated no exposure to any type of HRT. There would be an accurate assessment of exposure in never users; they would be comparable to the RCTs' placebo group in the sense that they have had no exposure to HRT. The problem lies with estimating exposure among the ever users. Among the ever users, there could be exposures to different types of EPRT, at different dosages and for different durations.

There were several methods that were used to assess the exposure of women to EPRT. Most of the studies (n = 12) simply asked the women to report on their use of EPRT in a questionnaire or interview. The researchers usually provided picture charts of

all prescribed brands of HRT to facilitate recall. The three remaining studies used pharmacy or doctors records to assess exposure. A review of these medical records would likely provide the most accurate assessment of EPRT exposure. However, according to Goodman et al[141], there is not much difference in the accuracy between any of these methods. In their study, they compared menopausal estrogen histories obtained through interview with those obtained from medical records. They concluded that women are able to recall HRT use with a high degree of accuracy and thus personal interviews are a valid method of obtaining accurate exposure information. The results reported by Paganini-Hill and Ross[142] also indicated that there is good agreement between medication information obtained through interview and that recorded in medical records. Therefore, a reasonable estimate of exposure assessment is provided through interviews. Furthermore, the comparability of the results obtained from two sets of methods indicates that all of the studies in this investigation are providing a similar assessment of exposure.

For those case control studies that relied on interviews to obtain information on EPRT exposure (n = 10), the cases would have reported their exposure information after learning of their breast cancer diagnosis. This could lead to recall bias[129], as women with breast cancer may recall circumstances that could have contributed to breast cancer, including HRT use, more accurately than controls. In this investigation, all studies reported on ever use of EPRT versus never use of HRT. Differential recall of postmenopausal hormone use is unlikely, as postmenopausal women, regardless of their disease status, are likely to remember whether or not they had ever taken supplemental hormones.

In the study comparing interview information with medical records mentioned above, Goodman et al[141] also compared the recall accuracy between breast cancer cases and two sets of control (hospital, neighbourhood) subjects. They found no differences in the recall accuracy among cases and controls in terms of ever versus never use of estrogens, duration of estrogen use and the age when the women initially used estrogen. Therefore, recall bias does not appear to be an issue for this investigation.

The random allocation component of a RCT ensures that the participants are randomly assigned (generally a blind allocation) to the treatment and placebo (or current gold standard treatment) groups, permitting investigators to control for all the disease risk factors, other than the risk of interest. As a result, the two groups contain a similar profile of participants. The similarity between groups is such that any outcome differences observed in the two groups can usually be attributed to the exposure of interest. In the nonrandomized cohort studies, women would not be randomly assigned to the EPRT/no HRT groups and thus there may be differences between exposed and unexposed groups such that any differences in outcomes cannot solely be attributed to exposure differences.

However, it is more than likely that the biases inherent in the nonrandomized studies would serve to reduce any EPRT-breast cancer association and thus would not explain the results observed here. The lifestyles of women that use any type of HRT have a tendency to differ from women not taking HRT. Women that decide to take HRT tend to be more health conscious; have no family history of breast cancer; have a negative mammogram before beginning any HRT use; and have higher education and socioeconomic levels[143]. Grodstein[144] describes a "healthy user" bias that exists among

HRT users. She reports that, in addition to the findings of Daly et al[143], women on HRT tend to be healthier, thinner, have less baseline heart disease and have different lifestyles than women who do not take HRT. This bias would be of particular importance when explaining the perceived cardiovascular benefits of HRT and why women on HRT have a lower incidence of heart disease or mortality. However, this "healthy user" bias would not be useful in explaining this investigation's finding that women on EPRT are at a higher risk of breast cancer. It is unlikely that their healthier lifestyle would place them at an increased risk of breast cancer[55]. In fact, recent studies have shown that women with healthier lifestyles defined by low fat diets[145], regular exercise[146] and low alcohol consumption[147] have a lower risk of breast cancer. Thus, at the outset, women on EPRT likely have a lowered risk of breast cancer. The fact that a significant association was found between breast cancer and EPRT use, despite this initial lowered risk, lends further credibility to the association.

In case control studies, one way to overcome the lack of random allocation is to control for the major risk factors in the design of the study or the analysis of the results. All of the case control studies incorporated in the meta-analysis did both. For instance, researchers matched controls with cases in terms of risk factors related to the outcome. One of the main risk factors for breast cancer is age[55]. All of the studies used in this investigation matched for age and thus cases and controls had a similar age profile. Other risk factors, such as parity, BMI and family history of breast cancer, were controlled in the analysis. Thus the studies reported adjusted odds ratios which controlled for these risk factors. One of the problems encountered during the analysis was the inconsistency in the choice of factors that were controlled in this manner. Greater consistency among

factors controlled in the analysis would permit more meaningful comparisons among studies. For three of the cohort studies, stratified analyses were conducted to control for risk factors. The risk estimates that were provided were thus adjusted for breast cancer risk factors. In the Laurtizen and Meier study[36], the authors claimed that the exposed and unexposed groups were of comparable age, body weight, parity and social status.

Even though the exposure assessment for nonrandomized studies would not be as accurate as it is for RCTs, the studies incorporated into this meta-analysis have relied on methodology that would ensure a reasonable exposure assessment. One factor that has often led to the dismissal of nonrandomized studies as valid sources of information has been the belief that, relative to RCTs, the results from nonrandomized studies report a stronger association between exposure and outcome. This was certainly the case for studies published in the 1970's and early 1980's. Miller et al[148] report that nonrandomized studies reported larger gains in treatment success than RCTs. Chalmers et al[149] also reported that 56% of the nonrandomized studies reported favourable treatment effects compared to only 30% of the RCTs. However, two recent studies[150;151] that compared RCTs with nonrandomized studies that were published since 1985, report the contrary.

Concato et al[150] compared the findings of 72 RCTs and 24 nonrandomized studies that examined the same five clinical topics. They evaluated studies that had been published between 1985 and 1998. They reported that the results between the two types of studies were "remarkably similar". They concluded that the results of well-designed nonrandomized studies did not overestimate the association as originally thought. While they did caution that nonrandomized studies that examine invasive therapies may be

subject to selection bias, they believed that studies involving drug therapies would be at less of a risk for this bias. Benson and Hartz[151] carried out a similar investigation and examined the results from 83 RCTs and 53 nonrandomized studies published between 1991 and 1995 that evaluated the same19 treatments. They reported little difference between the estimates of treatment effects from the two types of study designs.

In another recent report, Linde et al[152] assessed whether systematic reviews should include "nonrandomized and uncontrolled" studies. They evaluated the use of acupuncture as a treatment for chronic headaches and compared the findings of 24 RCTs with 35 non-RCTs consisting of 15 cohort studies, 10 case series and 10 cross sectional surveys. While they found that the non-RCTs reported findings similar to the RCTs, they reported that, in general, these non-RCTs provided little relevant information on key details. They also found that including nonrandomized studies in a systematic review takes a considerable amount of time because a broad search is needed to locate the studies, the extraction and analysis is more complex and quality assessment must be more comprehensive.

One of the problems with this study was that the majority of non-RCTs were case series and cross-sectional studies, both descriptive studies with considerable design limitations.[129] If their non-RCT evaluation was restricted to the 15 cohort studies, the majority of studies were similar in quality to the RCTs (met five or more of the quality criteria) and yielded similar response rates. A closer inspection of the studies included in their report revealed that almost 50% of the non-RCTs (17/35) were published before 1985, while only 25% of the RCTs (6/24) were published before then. Since study quality has generally been improving over time[153](section 5.4.2.1), the weaker quality of

the non-RCTs reported by Linde et al is not unexpected. They concluded that inclusion of good quality nonrandomized studies (in this case, the cohort studies) could provide a "useful extension to systematic reviews". Therefore, the evidence from these three reviews[150-152] indicate that well designed nonrandomized studies provide valid information and should not necessarily be excluded from studies of evidence-based medicine.

5.3 Quality assessment

5.3.1 Newcastle Ottawa Scale

The NOS, used to measure study quality, was a straightforward quality assessment tool. Its nine different items were easy to understand and score. Each study took, on average, approximately eight minutes to evaluate. This is considerably less than the 15 minutes required, on average, to assess the quality of a study using the Black and Down's scale (Appendix B). The NOS also had very good inter-rater reliability. The two reviewers awarded the same score to all but two of the 15 studies.

One problem with the NOS concerned the Comparability criterion. This criterion had limited discriminatory power because 14 of the 15 studies were awarded a score of two with the remaining study receiving a score of one. There was greater variability found in the scores for the other two criteria. Scores for the Selection criterion ranged from two to four and the Outcome/Exposure Assessment scores ranged from one to three. More work needs to be done on the Comparability criterion to improve the tool's ability to distinguish between poor and good quality studies. One suggestion is to differentiate between studies that controlled for important risk factors in the design of the study and

those that controlled for risk factors in the analysis. A higher score should be awarded to studies in the latter category. In addition, there were some studies that controlled for five or more risk factors and others that controlled for only two. Both sets of studies would receive the same maximum score of two. Once again, the scoring system needs to be adjusted to reflect differences in the control of risk factors, with studies receiving higher scores with increasing numbers of risk factors controlled.

5.3.2 Quality of studies

The Cochrane Collaboration identifies seven essential steps involved in the preparation and maintenance of a systematic review.[82] Step three involves the critical appraisal of studies. They provide three reasons for assessing the quality of studies: 1) to limit bias; 2) to provide insight into comparisons; and 3) to aid in data interpretation. Two of the nine criteria used by Oxman and Guyatt[64] to assess the scientific quality of systematic reviews and meta-analyses involve the assessment of study quality. Studies that report the criteria used for assessing study quality, providing those criteria are appropriate, receive higher a scientific quality rating compared to overviews that ignore study quality. Moher et al[154] also stressed the importance of assessing the quality of the research evidence. In a later paper, they report that in spite of this, few meta-analyses of RCTs assess the quality of their studies.[89] They found that ignoring study quality can influence the conclusions drawn from the analyses. Therefore, quality assessment is generally considered to be an important step in meta-analysis methodology, in particular for meta-analyses involving nonrandomized studies.[88]

Quality assessment is generally accepted as an important step in meta-analytic methodology[80;82;84;85]. However no one best scoring instrument has been identified[155]. NOS was chosen for this investigation because it was ranked among the top six tools for nonrandomized studies[91] and it provided similar assessment tools for both case control and cohort studies.

Juni et al[155] also reported that a clearly defined method of incorporating quality scores into a review is still under debate. There are several suggested methods for incorporating quality scores into a review. Detsky et al[88] reviewed four of these methods: 1) use a threshold score as a cut off point for inclusion/exclusion criterion (ie. studies below a certain score are excluded from analysis); 2) incorporate the quality score as a weight in statistical pooling; 3) construct a plot of effect size versus quality score; and 4) combine the study results sequentially, where the order of study inclusion is based on quality score.

When considering which of these four methods would be most appropriate for this investigation, one must consider the NOS and the nature of the scores obtained. The NOS has not defined a threshold score such that studies below a particular score are considered poor quality and above that score, good quality. There is also a certain degree of subjectivity involved in using any quality assessment tool. Therefore, using a subjective score, with an arbitrarily assigned threshold as part of the inclusion/exclusion criteria lacks rigour. Furthermore, decisions to reject a study based on quality score may vary according to the instrument used. Given the characteristics of the studies included in this meta-analysis (nonrandomized, smaller sample sizes), methods two and three would also be problematic. Detsky et al[88] argued that these statistical techniques would

lack power and empirical justification. While the authors concluded that there is "no right approach", they recommended the fourth method: the sequential combination of study results. This method was chosen to incorporate the NOS scores as it is best suited to this investigation (section 5.4.4).

There was limited variation in the quality scores of studies in this meta-analysis (scores ranged from five to eight out of nine). This would suggest that the evidence to judge the association between breast cancer and EPRT exposure was of good quality. One of the reasons for these relatively high scores may be due to the fact that all but three of the studies were published in the scientific literature and thus underwent the rigours of the peer review system. Because of the lack of variation in the quality assessment scores, it was difficult to discuss the implications of study quality on study outcome. The fact that all of the case control studies scored in the top third of the quality assessment scale suggests that study quality did not compromise the significance of these results. If the study scores fell in the bottom third of the scale, the significance of the results may be questionable. A discussion of the relationship between study quality and study results appears in the section 5.4.2.1.

5.4 Quantitative analysis

5.4.1 Ever users of estrogen progestin replacement therapy vs. never users of hormone replacement therapy

The "ever users" of EPRT from case control and cohort studies had a 40% increased risk of breast cancer compared to the "never users" of HRT. Statistical tests revealed that there was a significant association between breast cancer and EPRT

exposure among ever users. Among the six meta-analyses examining the relationship between breast cancer and ERT, four provided summary estimates of the association between breast cancer and ever use of ERT.[26;27;29;30] All four of these meta-analyses reported similar findings; i.e., there was no increased breast cancer risk associated with ever use of ERT. Their summary estimates ranged from 0.99 to 1.08. The differences in breast cancer risk between women on EPRT and ERT, as summarized in the meta-analyses, suggest that ever use of EPRT places postmenopausal women at a higher risk of breast cancer than ERT.

Colditz[156] argued that progesterone may act synergistically with estrogen in stimulating the proliferation of breast tissue. Progestins are reported to have a greater proliferative effect on either initiated cells or on the growth of cancer cells. Furthermore, breast cell proliferation in premenopausal women peaks during the menstrual cycle, at a time when the levels of progesterone are also at their peak.[157] Therefore, it is reasonable to argue that breast cell proliferation is greater in postmenopausal women on EPRT compared to women on ERT or never users of HRT.

This statement is supported by the recent findings of Hofseth et al.[60] They examined breast density and indices of breast cell proliferation among three groups of women: EPRT users; ERT users and never users. They found that while both groups of women on HRT had higher breast density than never users, the women on EPRT had the highest breast density and the highest breast cell proliferation index among the three groups. The site of breast cell proliferation also corresponded to the area where most breast cancers develop. High breast density, as revealed by mammographic screening, is a strong predictor of breast cancer, behind age and family history[61;62;158]. In a study of

ovariectomised macaques, Cline et al[159] found higher breast cell proliferation among the animals that had long term exposure to EPRT compared to those with long term exposure to ERT. Rutteman[160] also found that progesterone or synthetic progestins stimulated mammary tumour development in dogs, cats, rodents and monkeys. In rhesus monkeys, Tavassoli[161] reported a higher proportion of mammary gland carcinoma among those animals exposed to EPRT formulations compared to those exposed to ERT only. The results of these laboratory studies supports the epidemiological findings and helps to explain the increased breast cancer risk reported in this EPRT meta-analysis compared to earlier ERT studies.

The classification of "ever use" of EPRT is less than ideal. It encompasses all women who have ever used EPRT. It does not distinguish between women that have been on EPRT for ten years and those on EPRT for less than five years. The "ever use" label would also fail to distinguish between women who had been on EPRT for three years and then went off the therapy for two years from women who had been on EPRT continuously for five years.

In the majority of studies published before 1996, the focus of the study was on breast cancer and ERT use. Very few of the women in these earlier studies were on EPRT. It was only once researchers recognized the strong association between endometrial cancer and ERT use that EPRT became more widely used.[162] There were few details provided on EPRT use in these early studies because, according to the authors, there was insufficient data to carry out additional analyses. Therefore, it is difficult to gain further insight into "ever use" of EPRT. However, it is interesting to look at the two studies[115;118] that reported a decreased risk of breast cancer associated

with EPRT use. In the Palmer study, only three of the 18 women had been exposed to EPRT for longer than five years. In the Stanford study, only 11 of the 215 women had exposures longer than five years. Therefore, the reason that these studies reported a decreased risk may be due to the fact that most of the women had short term exposure to EPRT. Most of the studies published since 1996 reported "ever use" risk estimates that were significantly greater than one. In these later studies, there were more women on EPRT and their exposure to the therapy was longer than the exposure in the earlier studies.

The fact that this association was statistically significant and there was no significant heterogeneity found between the case control studies and the cohort studies indicates that the observed differences between studies were due primarily to chance and not some other factors. Subgroup analyses are generally recommended when the heterogeneity between studies is significant. However, one of the purposes of this review was to identify and examine differences between the studies to gain greater insight into the relationship between outcome and exposure (the analytic component[79]). Relying solely on the results from ever use of EPRT to assess the relationship between breast cancer and EPRT use is problematic due to the extensive range of exposures implied by "ever use".

Oxman and Guyatt[163] state that subgroup analyses permit the identification of characteristics that modify the effect of exposure on outcome. Thompson[164] argued that the meta-analysis should focus on understanding sources of heterogeneity through subgroup analyses. Therefore subgroup analyses were carried out in order to determine

the role that study quality, duration of EPRT use, recency of EPRT use, study location and EPRT formulation play in this relationship.

5.4.2 Subgroup analysis

5.4.2.1 Study quality

Petitti[80] identified study quality assessment as an essential component of a meta-analysis because it can enhance its validity. She argued that study quality provides an indication of study validity such that "studies of higher quality scores yield more valid information than studies with lower quality scores". When the risk estimates of studies with similar quality scores in this investigation were combined, an interesting pattern emerged. The analysis showed that the higher the study quality, the greater the risk estimate. The fact that the highest quality studies had the highest summary risk estimate further supports the conclusion that there is an increased risk of breast cancer associated with EPRT use. If it were the studies with lower quality scores that had the greater risk estimates, the association would be less convincing.

It is also interesting to note that the studies with the higher quality scores had a significant association between breast cancer risk and EPRT use, while the lower quality scores did not report this significant association. Closer examination of these studies indicate that study quality was related somewhat to the year the study was published. The studies with higher quality scores were published from the mid to late 1990's while the studies with lower quality scores were published in the early 1990's and 1980's. Emerson et al[153] also found that the quality level of research studies has increased at a rate of nine percent per decade (from 1960 to 1990).

One reason for this may be related to an improvement in the reporting of study details. Two recent papers [84;85] have provided comprehensive criteria for reporting meta-analyses of RCTs and nonrandomized studies. Moher et al[89] reported that methods to improve reporting quality have been endorsed by several journals and included in their instructions to authors. Therefore this trend of improving study quality will likely continue.

The improvement of study quality over time is also likely due to improvements in study design. Benson and Hartz[151] argued that more recent nonrandomized studies may be "methodologically superior" to earlier studies, due in part to more careful selection of participants. They argue that attention to study quality may account for improvements in both study design and reporting.

The results of this investigation are further supported by the results of Steinberg et al[28]. This paper was the only breast cancer/ERT meta-analysis that assessed the relationship between study quality and risk estimates. In their meta-analysis, they found that the studies with the highest quality were also the ones with the highest risk estimates. While they did not provide specific details about the scale that they used to assess quality, it did contain criteria similar to the NOS. For instance, studies received a higher quality score if the breast cancer cases were validated (NOS item one); if the cases and controls came from the same population (NOS item three) and if the analysis was adjusted for confounders (NOS item five). In addition, as in this current investigation, the higher quality studies also tended to be the more recent studies (published from 1980 to 1989), while the majority of studies with the lowest quality scores were published earlier (from 1977 to 1982). The similarity between the Steinberg's results and those of the current

meta-analysis provides additional support for the importance of study quality assessment and its relationship to study validity.

This subgroup analysis has also revealed that study quality can modify the relationship between exposure on outcome, such that studies of higher quality demonstrate a stronger association between breast cancer and EPRT use.

5.4.2.2 Duration of estrogen progestin replacement therapy

Duration of exposure is an important consideration when assessing the relationship between breast cancer and EPRT use. In spite of this, only five of the 15 studies[38;39;74;117;119] reported information about duration of EPRT exposure that was sufficient for inclusion into the meta-analysis. For instance, in their 1995 paper, Colditz et al[71] reported that they were unable to examine breast cancer risk associated with duration of EPRT use because the widespread use of progestins was a "recent phenomenon" in the United States. However among the five studies that did report duration of use data, the findings were consistent. The results from the case control studies indicated those women on EPRT for 10 years or longer had almost three times the risk of breast cancer compared to never users. Women on EPRT for less than five years had a 35% increased risk of breast cancer.

A within-study comparison of short term vs. long term duration was possible for four of the studies[38;39;74;117]. Among these studies, the findings were consistent: there was an increased risk of breast cancer associated with an increased duration of EPRT exposure. In two of these studies[39;117], there was a significant risk of breast cancer associated with long term use, but not when exposure was limited to short term use. It

was more difficult to make comparisons between studies, because different studies chose to assess differing lengths of EPRT duration. However, the same pattern emerged; an increased risk was associated with an increased duration.

These results are similar to the breast cancer/ERT meta-analyses. Three of these studies[28;30;31] reported a statistically significant risk of breast cancer associated with increasing duration of ERT exposure. While the increased risk was not as great as that noted for long term duration of EPRT use, there was still a 23-35% increased risk of breast cancer for women that had been on ERT for five years or longer. One meta-analysis[28] also reported that among studies with high quality scores, there was a statistically significant association between breast cancer risk and long term ERT use, but no such association existed among studies with lower quality scores. These findings further support the relationship between breast cancer and long term HRT use.

These results also lend further support to the notion that EPRT use poses a greater risk of breast cancer than ERT use. The breast cancer risk associated with ERT use for 15 years or longer was 1.30 (95% CI 1.2-1.6)[28]. In this investigation, the breast cancer risk associated with EPRT use for 10 years or longer was 2.94 (1.94-4.45). Ross et al[40] also found the breast cancer risk to be greater among long term EPRT users. In their study, the risk of breast cancer was expressed per five years of ERT or EPRT use. They report the OR for ERT as 1.06 (95% CI 0.97-1.15) and the OR for EPRT as 1.24 (95% CI 1.07-1.45), per five years of use. Since the risk associated with long term EPRT use was significant and larger than the risk for long term ERT use, it lends further support to the view that EPRT poses a higher breast cancer risk than ERT.

5.4.2.3 Current use of estrogen progestin replacement therapy

Women classified as current users of EPRT had a significantly higher risk of breast cancer compared to never users of HRT. There were six studies[39;71;74;116;117;127] in the meta-analysis that compared the risk between current and never users and all but one[117] reported a significant risk associated with current use. Women who are current users of EPRT may provide a more accurate assessment of their EPRT use in terms of duration or formulation than ever users who are no longer on the therapy. However, this is unlikely to affect the resulting risk estimates. In spite of the fact that the risk associated with current use was found to be higher than the risk of ever use, there were very few explanations that were offered to clarify these differences.

Colditz et al[71] proposed that the increased risk associated with current use may result from a potential detection bias resulting from differences in mammographic participation rates. There may be a higher frequency of screening mammography among women who are current users of EPRT compared to women who are past or never users. However, they found that the screening rate was only 14 percent higher among current users and eight percent higher among past users, relative to never users. Furthermore, most of the women in the study, regardless of their hormone use, had at least one mammogram. One study argued that multiple mammograms during a fixed period have "diminishing returns" and will not necessarily increase the probability of breast cancer detection.[165]

Joffe et al[165] recently re-examined the issue of mammographic frequency among postmenopausal women in the Nurse's Health Study and detection bias. They reported that any HRT use can produce symptoms that result in physician visits and perhaps more

mammography. While they found that women on any type of HRT have more mammograms, they argued that it cannot explain the association between HRT use and breast cancer. Another interesting point that they raised is that if this detection bias was ultimately responsible for this association, then the highest detection rates would be observed among women who had just started to use HRT, because this is the group that had the highest frequency of mammograms. The fact that the highest risk of breast cancer is seen among current *long term* users suggests that detection bias is not the only explanation. They state that their study provides further support of the relationship between breast cancer and HRT use.

5.4.2.4 Study location

Several reports have alluded to differences in HRT use between European and North American postmenopausal women[36;126;128]. For instance, Lauritzen and Meier[36]report that while their results were comparable to those from other European studies, they differed from North American studies.[37;130] They attribute this discrepancy to differences in doses of estrogens and the addition of progestins to HRT formulations. They argued that North American formulations used higher doses of estrogen (1.25 mg/day vs. 0.625 mg per day) and rarely added progestins, while the use of progestins was quite common in Europe. Several studies have reported a similar discrepancy between European and North American EPRT formulations [122;126].

While the summary risk estimate for European studies was higher, there was no significant difference noted between European and North American studies in this investigation. If the doses of estrogen were higher in North America, and it was the

estrogen in EPRT that increased the risk of breast cancer, then one would predict a higher risk among North American women. The results from this investigation do not support this prediction. This may be because there has been a change in the formulations used over the past 20 years. For instance, Colditz et al[71] reports that in the 1990's, use of EPRT was becoming more widespread among postmenopausal American women.

While Europe and North American physicians may be becoming more consistent in their prescribing practices, a difference still remains in the types of progestins they prescribe. The effects of these different types of progestins are described in the next section.

5.4.2.5 Testosterone vs. progesterone derived estrogen progestin replacement therapy

Among those studies that described the estrogen and progestin components of EPRT, all reported the same type of estrogen in the formulation: conjugated equine estrogens. However, there were differences in the types of progestin used and these differences were based on study location. A number of studies[128;132] have reported that in Europe, the most commonly used progestin formulation was testosterone derived, while in North America, it was progesterone derived. Magnussen et al[74] reported that 90% of the women on EPRT were using a testosterone derived progestin. Persson et al[38] found that 45% of the Swedish women in their cohort also used a testosterone derived progestin (Levonorgestrel or Norethisterone acetate), while 55% used a progesterone derived formulation (medroxyprogesterone acetate). Medroxyprogesterone acetate was also the

type of progesterone derived progestin commonly used in North American EPRT formulations [115;116].

There was a significantly higher breast cancer risk for women on EPRT with testosterone derived progestins. Of the studies that provided details concerning EPRT formulation, there was no significant association between breast cancer risk and exposure to progesterone derived EPRT. This may be a result of the fact that these different types of progestins have different effects on breast tissue.

Santen et al [72] argued that progestins differ in structure and function and thus can behave differently, depending on the tissue they are found in. So, while progestins counteract the proliferative effects that estrogens have on the endometrial tissue, they could exert proliferative or suppressive effects on breast tissue. The study by Hosfeth et al [60] looked exclusively at a progesterone derived progestin, and found that it had a proliferative effect on breast tissue, particularly in those areas of the breast where most cancers develop. Campagnoli et al [166] examined the effect of testosterone derived progestins on breast cancer cell lines. They reported that the testosterone derived progestins have strong androgenic effects that oppose the estrogenic effects. They argued that the action of the testosterone derived progestins lead to increased stimulation of the breast. Therefore, it posed a higher risk than estrogen alone or progesterone derived progestins that only have slightly androgenic properties.

The results from this investigation suggest that testosterone derived progestins present a greater breast cancer risk than progesterone derived progestins. There are several studies that support the presumed proliferative effects of a testosterone derived compound. For instance, Ewertz[113] reported on the risk of breast cancer in relation to

different types of sex hormones used. Of the four different types of formulations (ERT, progestin only, EPRT, and estrogen/androgen), she found the highest breast cancer risk was for the estrogen/androgen combination (OR = 2.31; 95% CI 1.37-3.88). The androgen used in this formulation consisted of 50-100 mg of testosterone. Her study supports the contention that an androgenic compound increases the risk of breast cancer and concurs with the laboratory studies described above.

Magnussen et al [74] compared the risk associated with testosterone and progesterone derived progestins and reported a significant increased risk for the testosterone derived formulation, but no significant risk associated with the progesterone derived formulation. Given the notable increased risk associated with androgen based compounds, caution should be taken before prescribing EPRT with a testosterone derived progestin, particularly for long term use.

5.4.3 Sensitivity analysis

The results of the sensitivity analyses for case control and cohort studies indicated that including studies that were excluded for a variety of reasons (unadjusted risk estimates, no peer review, did not meet inclusion/exclusion criteria, use of hospital controls) had little effect on the results. When this occurs, Clarke and Oxman [82] stated that "it strengthens the confidence that can be placed in these results".

In the case control sensitivity analysis, the inclusion of two studies[40;122] that were initially excluded for methodological flaws did not influence the summary estimate or the results of the statistical tests. Including the unpublished studies[116;117] and those that used hospitals as their source of controls[120;167] did not alter the results either.

The same holds true for the cohort/RCT sensitivity analysis. The inclusion of one or both of the RCTs[19;133] did not influence either the summary estimate or statistical tests, nor did the inclusion of the cohort study with an unadjusted risk estimate[37] or the one study that was not peer reviewed[36].

The results of these sensitivity analyses indicate that the association between increased breast cancer risk and EPRT exposure is robust. It is also interesting to note the almost identical results obtained for the case control and cohort sensitivity analyses. The risk estimates in all steps of the case control analysis ranged from 1.36 to 1.42 and from 1.36 to 1.41 in the cohort analysis. Their confidence intervals were also very similar. The fact that the results from these two types of studies are in close agreement provides additional support for the validity of these results.

5.4.4 Cumulative meta-analyses

Traditionally, cumulative meta-analyses have been used to assess the robustness of the risk estimate over time. With these chronological meta-analyses, the risk estimate shifted slightly over time, to indicate that the risk was higher in the more recent studies. The results from the cumulative meta-analyses in this investigation indicate that the breast cancer risk associated with EPRT use were slightly modified over time and with declining study quality.

The shift in summary estimates from no risk to increasing risk in the chronological cumulative meta-analysis indicate that over time the breast cancer risk associated with EPRT exposure increased. This is likely due to the fact that the more recently published studies involve postmenopausal women who have been on EPRT

longer. In the 1970's, most of the prescriptions for progestins were used alone and only 18% of those prescriptions were for treatment of menopausal symptoms. The combined EPRT formulation became popular in North America in the early 1980's. By 1986, over 60% of the progestin prescriptions were in EPRT and used for the treatment of menopausal symptoms[4]. Colditz et al[71] charted changes in EPRT use and reported that the number of nurses in their study on EPRT gradually increased from a low of 18% in 1986 to 30% four years later. The more recent studies would include postmenopausal women that have been on EPRT for a longer period of time. Therefore, the fact that the cumulative meta-analyses showed an increase in the breast cancer risk with the sequential addition of more recent studies is consistent with the findings reported for the relationship between breast cancer risk and long term EPRT use.

The results of the cumulative meta-analyses related to study quality indicated that the breast cancer risk associated with EPRT use declined slightly with declining study quality. The results of this cumulative meta-analysis are consistent with the findings reported for the relationship between breast cancer risk and study quality discussed in section 5.4.2.1. It further demonstrates that as study quality declines, the breast cancer risk associated with EPRT use also declines. If the trend was reversed (i.e. the risk of breast cancer increased with decreasing study quality), then the strength of association between breast cancer risk and EPRT use would be questionable. It would indicate that the relatively poorer quality studies show a greater risk and the higher quality studies do not. Given that quality is often equated with validity[80], the results of the cumulative meta-analyses of study quality suggest that it is the more valid studies that have the

higher risk estimates. This provides further support to the robustness of the association between breast cancer and EPRT use.

Pettiti[80] argued that study quality often reflects the era in which the studies were conducted. Apart from the Kirsh study (an unpublished MSc thesis)[117], the higher quality studies were published after 1995 and the relatively lower quality studies were published during or before 1995. Therefore, these differences in study quality may reflect improvements in the design and implementation of nonrandomized studies.[150;151] Petitti also suggested that studies that have been published more recently are rated as higher in quality because reporting standards have improved. However, the differences in quality scoring between the studies had little to do with poor reporting. Rather, the differences in scores were due to nonrandom samples of women, insufficient adjustment for risk factors and poor response rate.

5.4.5 Publication bias

Publication bias is said to occur when there are more studies with statistically significant results submitted and published compared to studies showing statistically non-significant results. This would mean that it would be easier to locate studies with statistically significant results and these would be the ones included in the meta-analysis. This would bias the summary estimates and lead to conclusions that were based on only a subset of the evidence. Pettiti [80] stated that the issue of publication bias will be resolved when good quality studies are submitted and accepted regardless of their findings. For the association between breast cancer and EPRT use, one would not predict an inherent

publication bias for studies demonstrating an increased or decreased risk with EPRT use or no risk.

There are several reasons why it would be unlikely to find a publishing bias for studies examining the relationship between breast cancer and EPRT. First of all, there have been relatively few studies that have examined the relationship between breast cancer and EPRT use. EPRT is becoming a more commonly used HRT formulation[71] and thus more and more individuals are interested in its associated health risks. It is unlikely that an author would choose not to publish their findings because they found either a decreased or an increased risk. In this meta-analysis, three of the published studies[36;115;118] showed a decreased risk of breast cancer associated with EPRT use and the remaining 12 showed an increased risk. However, the results of only five of these studies were statistically significant. The most recent meta-analysis of the breast cancer/ERT relationship[30] examined 29 different studies. Twelve of these studies reported a reduced risk, 11 reported an increased risk and three reported no risk. Only three of these studies had statistically significant results. This supports the position that there was no publication bias associated with studies examining the relationship between breast cancer and HRT use.

There does not appear to be any suitable statistical approaches for assessing the presence of publication bias in nonrandomized studies. While there have been other techniques developed to assess publication bias, most pertain to RCTs.[168] One of the more popular methods used to assess publication bias, described as a "quasi-statistical" approach[80], involves the construction of funnel plots. This technique involves plotting some measure of sample size against relative measures of exposure or treatment effects

(in this case, odds ratios and relative risks) and evaluating whether the resulting plot is shaped like an inverted funnel. This is a rather subjective assessment and furthermore, there has been no evaluation of the funnel plot's sensitivity in detecting publication bias.[80]

Since a recent study[104] recommended using log odds ratio as the measure for treatment effect and standard error as the measure of study size, these were chosen as the specific x and y values. While apparently straightforward, there was a problem encountered in this study when deciding on which values to plot.

In this investigation, all of the studies reported findings on the relationship between breast cancer and exposure to both ERT and EPRT. In all cases, there were more women on ERT than on EPRT and the major focus of the paper tended to be on the relationship between breast cancer and ERT exposure. Because it was difficult to decide whether to plot the EPRT data (which was the focus of this investigation) or the ERT data (which in most cases was the focus of the publication), both sets of data were graphed.

The shapes of the two plots somewhat resembled the ideal inverted funnel and this would indicate that there was no apparent publication bias. The only gap that appeared in the plots corresponded to studies of intermediate size and a significant treatment effect. It is unlikely that there would be a bias against publishing studies with this profile. One would expect a publication bias against small studies that demonstrated little if any treatment effect. There were at least two studies in each of the funnel plots that fit this description. Therefore, this rather subjective evaluation of the two funnel plots indicates that both approximate the shape of an inverted funnel and thus there was no publication bias.

There were several strategies employed in this investigation to reduce publication bias. For instance, Egger et al[169] argued that English language bias could be introduced into systematic reviews if only English language journals were searched. They reported that the international papers with significant findings may choose to publish their findings in English language journals, while ones with nonsignificant findings would be relegated to local non-English journals. Therefore, if a literature search was restricted to English only journals, then a bias towards significant effects may result. When developing the search strategy for the electronic databases, no language restrictions were used. The majority of the 28 non-English articles that were revealed during the search were written in French, with three articles written in Italian and five in German. Fortunately, English abstracts accompanied each of these articles. If there was still some doubt of the suitability of a particular article after reviewing the abstracts, the article was retrieved. Only one non-English article was chosen for the final analysis[108], but upon closer inspection, its English language duplicate had already been selected for inclusion.[109]

One could argue that there was a language bias in this investigation because all of the studies were published in English. However, based on a study by Moher et al[170], this would not necessarily imply that the summary estimates determined in this investigation were biased. They examined the ramifications of restricting the articles used in a meta-analysis to English only. They reported little difference in the summary estimates between meta-analyses that included studies published in different languages and those in English only. They found no evidence that using English only articles leads to a bias in the summary estimates.

Another area that may contribute to bias involves the inclusion of grey or unpublished literature in the meta-analysis. There is considerable controversy surrounding the issue of whether or not to incorporate unpublished work into the analysis. Easterbrook et al[171] recommended the inclusion of grey literature, otherwise a cautious assessment of the results is required. They reported that nonrandomized studies were at a greater risk of publication bias than RCTs, but did not explain their reasoning. The fact that the published studies included in this investigation had significant and non-significant findings does not support their claim.

Dickersin[135] stated that while inclusion of the grey literature may provide pertinent information, she cautioned that it is not necessarily a cure for publication bias. Unlike searching the published literature, there is no unbiased approach to gain access to the grey literature. However Smith[171] argued that failing to include grey literature could lead to "misleading generalizations". McAuley et al[172] reported that excluding grey literature could contribute to bias in the findings of a meta-analysis. They recommended that all pertinent studies, published or not, should be included in the analysis. The Oxman and Guyatt quality assessment scale for meta-analysis characterizes a comprehensive search as one that encompasses both electronic databases and unpublished research.[64] However, among the breast cancer/ERT meta-analyses, none of the studies used unpublished literature. In fact, one of the meta-analyses[28] completely dismissed the inclusion of grey literature. Steinberg et al argued that unpublished data could be unreliable and believed that most credible work would be published anyway, thus alleviating the need to use unpublished research. They also stated that because many of

the published studies used in their meta-analysis had negative and null results, the risk of publication bias is minimal.

In this investigation, data from two unpublished studies were included in the analysis. Both were from graduate student theses.[116;117] One study[171] reported that that the average effect found in published studies is always greater than the effect reported in theses. However, the data from the theses were quite similar to the data obtained from published studies. In fact, the results of the sensitivity analysis revealed that the removal of these two studies did not affect the summary estimate. Therefore, the results from this investigation contradict the findings of Smith[171] and suggest that the significant association between breast cancer and EPRT remains robust even when the grey literature is excluded.

5.5 Conclusions and Future Directions

EPRT is one of the most popular HRT formulations currently prescribed to postmenopausal women worldwide. The results of this investigation indicate that there is an increased breast cancer risk associated with EPRT use among postmenopausal women, and this risk is considerably greater than the risk that has been previously reported for ERT use. This was particularly evident with long term EPRT use and when the EPRT formulations were testosterone derived. While EPRT may be a reasonable therapy for the management of short term menopausal symptoms, caution should be taken when it is being used to treat long term symptoms, such as osteoporosis.

Women who chose HRT because of its advantages in treating the short and long symptoms of menopause must continue to take supplemental estrogen in order to benefit from the reduction of bone loss and fracture risk associated with ERT or EPRT use.[173] Schneider et al[174] reported that the bone density of former ERT users that had been off ERT for ten years had the same bone density and risk of fracture as never users. This presents a dilemma for the postmenopausal woman and her physician since long term use of both ERT and EPRT, while necessary to prevent osteoporosis, also places these women at an increased risk of breast cancer.

There is a similar dilemma associated with testosterone derived EPRT. While the breast cancer risk associated with testosterone derived EPRT was found to be significantly higher, there are also benefits that have been linked to its use. For instance, Ventura et al [175] reported that postmenopausal women on this particular formulation may have a lower risk of cardiovascular disease. They found that patients receiving this formulation had reduced levels of homocysteine compared to women on placebo. High levels of homocysteine have been linked to atherosclerosis. Sarrel[176] reported that HRT with estrogen plus testosterone provided greater improvement in psychological and sexual symptoms associated with menopause (depression, reduced libido) than estrogen alone. Stroupe[177] found that this formulation may also increase bone density more than ERT. While none of these studies drew comparisons with progesterone derived EPRT, they certainly suggest that testosterone derived EPRT may be more effective at treating postmenopausal symptoms.

Traditionally, HRT was also touted for its benefits in reducing heart disease among postmenopausal women. The results of more than 40 observational studies

indicated that postmenopausal women taking supplemental estrogen had a reduced risk of heart disease.[18] However, the results of several RCTs have since contradicted these findings.[19;48]

In the HERS (Heart and Estrogen/progestin Replacement Study) Trial, the primary aim was to examine the effect of EPRT on the risk of myocardial infarction or death from coronary heart disease among postmenopausal women with established coronary disease.[19] In the first year of the study, they reported a 52% increase in cardiovascular events among women on EPRT. After four years of follow-up, they found no differences in the rates of myocardial infarction or coronary deaths between the treatment and placebo groups. The Estrogen Replacement and Atherosclerosis (ERA) Trial[48] reported that HRT was ineffective in slowing the progression of atherosclerosis among postmenopausal women at increased risk of cardiovascular disease, after one year of EPRT use. As a result of these findings, the American Heart Association recently published an advisory[178] which dismissed the use of HRT for secondary prevention of cardiovascular disease among postmenopausal women and indicated that evidence supporting the use of HRT for the "sole purpose of primary prevention of cardiovascular disease" is lacking.

The findings of these studies suggest that, while EPRT is effective at osteoporosis prevention, given the other health risks associated with its use, alternative therapies may be more appropriate. One such alternative is Raloxifene. Raloxifene is one of a series of SERM (selective estrogen receptor modulators) compounds. Raloxifene has been approved by the Food and Drug Administration (FDA) in the United States for the prevention and treatment of osteoporosis in postmenopausal women. While it increases

bone density, it is not as effective as estrogen is in reducing bone loss. However, it does not appear to increase the risk of breast cancer. In fact, Raloxifene is believed to reduce the incidence of invasive breast cancer in postmenopausal women with osteoporosis.[179] In a related study, Freedman et al [63] examined changes in mammographic density among healthy postmenopausal women on ERT, Raloxifene and placebo for a two year period. They found that breast density increased only in women on ERT, and actually reported a decrease in breast density among women on Raloxifene and placebo. While it has no effect on serum levels of triglycerides and HDL cholesterol, it does decrease levels of total and LDL cholesterol and thus may provide some cardiovascular benefits.

Tamoxifen is another SERM that has been FDA approved for the treatment of breast cancer and to reduce the incidence of breast cancer among high risk women. It also has been reported to reduce bone loss in postmenopausal women and lower serum levels of total and LDL cholesterol.

While SERMs have several advantages over HRT, they also have several deficiencies that make then unsuitable as a HRT replacement. For instance, Tamoxifen is associated with a fourfold increased risk of endometrial cancer.[105] In addition, there is a risk of venous thromboembolism associated with both Taxomifen and Raloxifene. SERMs are also ineffective at treating short term menopausal symptoms, and some women even experience a worsening of vasomotor symptoms.[180] Therefore, while SERMs initially appeared promising, more research on suitable alternatives for long term HRT is needed.

The best treatment for postmenopausal symptoms is one that takes into account the risks and the benefits of the therapy, and how it relates to the particular health profile

of the postmenopausal woman. If a woman is taking EPRT to treat the short term symptoms of menopause, then short term use of EPRT may prove to be a better alternative to a SERM. However, if long term therapy is required, alternative therapies will need to be explored. Over the next decade, the results of the Women's Intervention Study of long Duration Oestrogen after Menopause (WISDOM)[181], the Women's Health Initiative[8] and the Million Women Study[182] will provide additional insight into the risks and benefits of long term HRT use.

The actual formulation of the progestin also needs to be carefully considered. Future studies examining the relationship between breast cancer risk and EPRT use need to obtain a more detailed assessment of the EPRT formulation used so the risks and benefits of different progestins can also be better understood. Further research into women's attitudes towards taking HRT in light of these risks and benefits would also be useful in the guiding the decisions of postmenopausal women.

Literature cited

1. Statistics Canada. Annual Demographics Statistics Catalogue. 2000. Ottawa, Canada, 91-213.

2. Paganini-Hill A. Estrogen replacement therapy in the elderly. *Zentralblatt fur Gynakologie* 1996;**118**:255-61.

3. Statistics Canada. Report on the Demographic Situation in Canada 1997. Current Demographic Analysis. (Catalogue no.91-209-XPE,). 1998. Ottawa, Canada.

4. U.S.Congress. The menopause, hormone therapy and women's health. Office of Technology Assessment OTA-BP-BA-88, 1-119. 1992.

5. Boston Collaborative Drug Surveillance Program. Surgically confirmed gallbladder disease, venous thromboembolism and breast tumors in relation to postmenopausal estrogen therapy. *New Engl J Med* 1974;**290**:15-9.

6. Belchetz PE. Hormonal treatment of postmenopausal women. *New Engl J Med* 1994;**330**:1062-71.

7. Dennerstein L, Dudley EC, Hopper JL, Guthrie JR, Burger HG. A prospective population based study of menopausal symptoms. *Obstet Gynecol* 2000;**96**:351-8.

8. Barrett-Connor E,.Stuenkel CA. Hormone replacement therapy--risks and benefits. *Int J Epidemiol* 2001;**30**:423-6.

9. Lobo RA. Cardiovascular implications of estrogen progestin therapy. *Obstet Gynecol* 1990;**75**:18-25.

10. Clemens JD, Chuong JH, Feinstein AR. The BCG controversy: a methodological and statistical reappraisal. *JAMA* 1983;**249**:2362-9.

11. Greendale GA, Reboussin BA, Hogan P, Barnabei VM, Shumaker S, Johnson S *et al.* Symptom relief and side effects of postmenopausal hormones: results from the postmenopausal estrogen/progestin interventions trial. *Obstet Gynecol* 1998;**92**:982-8.

12. Levin RM. Osteoporosis: prevention is key to management. *Geriatrics* 1993;**48**:18
24.

13. Cummings SR, Kelsey JL, Nevitt MC. Epidemiology of osteoporosis and osteoporotic fractures. *Epidemiol Rev* 1985;**7**:178-208.

14. Writing group for the PEPI trial. Effects of hormone therapy on bone mineral density: results from the postmenopausal estrogen/progestin interventions (PEPI) trial. *JAMA* 1996; **276**:1389-96.

15. Ross RK, Mack TM, Paganini-Hill A. Menopausal estrogen therapy and protection from death from ischaemic heart disease. *Lancet* 1981;**1981**:858-60.

16. Stampfer MJ, Willett WC, Colditz GA. A prospective study of postmenopausal estrogen therapy and coronary artery disease. *New Engl J Med* 1985;**313**:1104-9.

17. Henderson BE, Paganini-Hill A, Ross RK. Estrogen replacement therapy and protection from acute myocardial infarction. *Am J Obstet Gynecol* 1988;**159**:312-7.

18. Manson JE, Martin KA. Postmenopausal Hormone Replacement Therapy. *New Engl J Med* 2001;**345**:34-40.

19. Hulley S, Grady D, Bush T, Furberg C, Herrington D, Riggs B *et al.* Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS) Research Group. *JAMA* 1998;**280**:605-13.

20. Nanda K, Bastian LA, Hasselblad V, Simel DL. Hormone replacement therapy and the risk of colorectal cancer: a meta-analysis. *Obstet Gynecol* 1999;**93**:880-8.

21. Yaffe K, Sawaya G, Lieberburg I, Grady D. Estrogen therapy in postmenopausal women: effects on cognitive function and dementia. *JAMA* 1998;**279**:688-95.

22. MacLennan SC, MacLennan AH, Ryan P. Colorectal cancer and oestrogen replacement therapy: a meta-analysis of epidemiological studies. *Med J Aust* 1995;**162**:491-493.

23. Mulnard RA, Cotman CW, Kawas C, van Dyck CH, Sano M, et al. Estrogen replacement therapy for treatment of mild to moderate Alzheimer Disease: a randomized controlled trial. *JAMA* 2000;**283**:1007-15.

24. Weiss NS, Ure CL, Ballard JH. Decreased risk of fractures of the hip and lower forearm with postmenopausal use of estrogen. *New Engl J Med* 1980;**303**:1195-8.

25. Cauley JA, Seeley DG, Ensrud K, Ettinger B, Black D, Cummings SR. Estrogen replacement therapy and fractures in older women. *Ann Intern Med* 1995;**122**:9-16.

26. Armstrong B. Oestrogen therapy after the menopause--boon or bane? *Med J Aust* 1988;**148**:213-4.

27. Dupont WD,.Page DL. Menopausal estrogen replacement therapy and breast cancer. *Archiv Int Med* 1991;**151**:67-72.

28. Steinberg K, Thacker S, Smith J. A meta-analysis of the effect of estrogen replacement therapy on the risk of breast cancer. *JAMA* 1991;**265**:1985-90.

29. Sillero-Arenas M, Delgado-Rodriguez M, Rodigues-Canteras R, Burger CW. Menopausal hormone replacement therapy and breast cancer: a meta-analysis. *Obstet Gynecol* 1992;**79**:286-94.

30. Colditz GA, Egan KM, Stampfer MJ. Hormone replacement therapy and risk of breast cancer: results from epidemiological studies. *Am J Obstet Gynecol* 1993;**168**:1473-80.

31. Beral V, Bull D, Doll F, Key T, Peto R, Reeves G. Breast cancer and hormone replacement therapy: collaborative analysis of data from 51 epidemiological studies of 52 705 women with bresat cancer and 108 411 women without breast cancer. *Lancet* 1997;**350**:1047-59.

32. National Cancer Institute of Canada. Canadian Cancer Statistics 2001. Toronto, Canada.

33. On L, Semenciw RM, Mao Y. Orius software: calculation of rates and epidemiologic indicators, and preparation of graphic output. *Chron Dis Canada* 2000;**21**:134-6.

34. Grady D, Gebretsadik T, Kerlikowske K, Ernster V, Petitti DB. Hormone replacement therapy and endometrial cancer risk: a meta-analysis. *Obstet Gynecol* 1995;**85**:304-13.

35. Nachtigall LE, Nachtigall RH, Nachtigall RD, Beckman EM. Estrogen replacement therapy. II. A prospective study in the relationship to carcinoma and cardiovascular and metabolic problems. *Obstet Gynecol* 1979;**554**:74-9.

36. Lauritzen C and Meier F. Risks of endometrial and mammary cancer morbidity and mortality in long-term estrogen treatment. van Herendael H, Riphagen F, and Goessens L. The climacteric--an update. 207-216. 1984. Lancaster England, MTP Press Ltd.

37. Gambrell RD Jr. Cancer and the use of estrogens. *Int J Fertility* 1986;**31**:112-22.

38. Persson I, Weiderpass E, Bergkvist L, Bergstrom R, Schairer C. Risks of breast and endometrial cancer after estrogen and estrogen-progestin replacement. *Cancer Causes & Control* 1999;**10**:253-60.

39. Schairer C, Lubin J, Troisi R, Sturgeon S, Brinton LA, Hoover R. Menopausal estrogen and estrogen-progestin replacement therapy and breast cancer risk. *JAMA* 2000;**283**:485-91.

40. Ross RK, Paganini-Hill A, Wan PC, Pike MC. Effect of hormone replacement therapy on breast cancer risk: estrogen versus estrogen plus progestin. *J Natl Cancer Instit* 2000;**92**:328-32.

41. Campbell NA. Biology. Fourth Edition. 1996. New York, The Benjamin/Cummings Publishing Company, Inc.

42. Casper RF, Morin-Gonthier M, Carter CJ, Edouard L, and Lefebvre Y. Menopause. 1-99. 1995. Canada, Minister of Supply and Services.

43. Riggs B, Wahner HW, Melton LJ. Rates of bone loss in the appendicular and axial skeletons of women. *J Clin Invest* 1986; **77**:1487-91.

44. Bush T, Cowan LD, Barrett-Connor E. Estrogen use and all cause mortality: preliminary results from the lipid research clinics program follow-up study. *JAMA* 1983;**249** :903-6.

45. Walter S, Wolf H, Barlebo H. Urinary incontinence in post-menopausal women. *Urol Int* 1978; **33**:135-43.

46. Torgerson DJ, Bell-Syer SEM. Hormone replacement therapy and prevention of nonvertebral fractures: a meta-analysis of randomized trials. *JAMA* 2001; **285**:2891-7.

47. Barrett-Connor E. Hormone replacement therapy. *BMJ* 1998; **317**:457-61.

48. Herrington D, Reboussin BA, Brosnihan KB. Effects of estrogen replacement on the progression of coronary artery atherosclerosis. *New Engl J Med* 2000; **343**:522-9.

49. Kimura D. Estrogen replacement therapy may protect against intellectual decline in postmenopausal women. *Horm Behav* 1995;**29**:312-21.

50. Kampen DL,.Sherwin BB. Estrogen use and verbal memory in healthy postmenopausal women. *Obstet Gynecol* 1994;**83**:979-83.

51. Leblanc ES, Janowsky J, Benjamin KS, Chan MS, Nelson HD. Hormone Replacement Therapy and Cognition: Systematic Review and Meta-analysis. *JAMA* 2001;**285**:1489-99.

52. Fernandez E, Franceschi S, La Vecchia C. Colorectal cancer and hormone replacement therapy. A review of epidemiologic studies. *J Br Meno Soc* 2000;**6**:8-14.

53. Koukoulis G. Hormone replacement therapy and breast cancer risk. *Ann New York Acad Sci* 2000;**132**:422-8.

54. Murray CJL,.Lopez AD. Mortality by cause for eight regions of the world: global burden of disease study. *Lancet* 1997;**349**:1269-76.

55. Key TJ, Verkasalo PK, Banks E. Epidemiology of breast cancer. *Lancet Oncology* 2001;**2** :133-40.

56. Kelsey JL. Breast cancer epidemiology: summary and future directions. *Epidemiol Rev* 1993;**15**:256-63.

57. Wolff MS,.Zeleniuch-Jacquotte A DN. Risk of breast cancer and organochlorine exposure. *Cancer Epidemiology, Biomarkers & Prevention* 2000;**9**:271-7.

58. Zumoff B. Does postmenopausal estrogen administration increase the risk of breast cancer? Contributions of animal, biochemical, and clinical investigative studies to a resolution of the controversy. *Proc Soc Exp Biol Med* 1998;**217**:30-7.

59. Pike MC, Spicer DV, Dahmoush L, Press MF. Estrogens, progestogens, normal breast cell proliferation and breast cancer risk. *Epidemiol Rev* 1993;**15**:17-35.

60. Hofseth LJ, Raafat AM, Osuch JR, Pathak DR, Slomski CA, Haslam SZ. Hormone replacement therapy with estrogen or estrogen plus medroxyprogesterone acetate is associated with increased epithelial proliferation in the normal postemenopausal breast. *J Clin Endocrin Metabol* 1999;**84**:4559-65.

61. Oza AM, Boyd NF. Mammographic parenchymal patterns: a marker of breast cancer risk. *Epidemiol Rev* 1993;**15**:196-208.

62. Boyd NF, Byng JW, Jong RA, Fishell EK, Little LE, Miller AB. Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening study. *J Natl Cancer Instit* 1995;**87**:670-5.

63. Freedman M, San Martin J, O'Gorman J, Eckert S, Lippman M, Shih-Chung B *et al.* Digitized mammography: a clinical trial of postmenopausal women randomly assigned to receive raloxifene, estrogen or placebo. *J Natl Cancer Instit* 2001; **93**:51-6.

64. Oxman AD,.Guyatt GH. Validation of an index of the quality of review articles. *J Clin Epidemiol* 1991;**44**:1271-8.

65. Weiss NS, Szekely D, English D, Schweid A. Endometrial cancer in relation to patterns of menopausal estrogen use. *JAMA* 1979;**242**:261-4.

66. Hulka BS. Replacement estrogens and risk of gynecologic cancers and breast cancer. *Cancer* 1987;**60**:1960-4.

67. Kurman RJ, Kaminski PF, Norris HJ. The behavior of endometrial hyperplasia. A long-term study of 'untreated' hyperplasia in 170 patients. *Cancer* 1985;**56**:403-12.

68. Whitehead MI, Lane G, Dyer G, Townsend PT, Collins WP, King RJ. Actions of progestins on the morphology and biochemistry of hte endometrium of postmenopausal women receiving low-dose estrogen therapy. *Br J Obstet Gynecol* 1981;**88**:914-8.

69. Persson I, Adami HO, Bergkvist L. Risk of endometrial cancer after treatment with oestrogens alone or in conjunction with progestogens: results of a prospective study. *BMJ* 1989;**298**:147-51.

70. Pike MC and Ross RK. Progestins and menopause: epidemiological studies of risks of endometrial and breast cancer. Steroids 65, 659-664. 2000.

71. Colditz GA, Hankinson SE, Hunter DJ, Willett WC, Manson JE, Stampfer MJ *et al.* The use of estrogens and progestins and the risk of breast cancer in postmenopausal women. *New Engl J Med* 1995;**332**:1589-93.

72. Santen RJ, Pinkerton J, McCartney C, Petroni GR. Risk of breast cancer with progestins in combination with estrogen as hormone replacement therapy. *J of Clin Endocrinol Metabol* 2001;**86**:16-23.

73. Greendale GA, Reboussin BA, Sie A, Singh HR, Olson LK, Gatewood O. Effects of estrogen and estrogen-progestin on mammographic paarenchymal density. *Ann Intern Med* 1999;**130**:262-9.

74. Magnusson C, Baron J, Correia N., Bergstrom R, Adami HO, Persson I. Breast cancer risk following long term oestrogen- and oestrogen-progestin-replacement therapy. *Int J Cancer* 1999;**81**:339-44.

75. Dixon JM. Hormone replacement therapy and the breast. We should worry about the increase in the risk of breast cancer. *BMJ* 2001;**323**:1381-2.

76. Klassen TP, Jadad AR, Moher D. Guides for reading and interpreting systematic reviews. *Archiv Ped Adolesc Med* 1998;**152**:700-4.

77. Cook DB, Sackett DL, Spitzer WO. Methodologic guidelines for systematic review of randomized controlled trials in health care form the Potsdam consultation on meta-analysis. *J Clin Epidemiol* 1995;**49**:235-43.

78. Egger M,.Smith GD. Meta-analysis: potential and promise. *BMJ* 1997;**315**:1371-4.

79. Greenland S. Meta-analysis. Rothman KJ and Greenland S. Modern Epidemiology. 643-674. 1998. New York, Lippincott Williams &Wilkins.

80. Petitti DB. Meta-analysis, Decision Analysis, and Cost effectiveness analysis. 2000. Oxford, Oxford University Press.

81. Greenland S. Can meta-analysis be salvaged? *Am J Epidemiol* 1994;**140**:783-7.

82. Clarke M and Oxman AD. Cochrane Reviewers Handbook 4.1.2. The Cochrane Library. (2). 2001. Oxford, Update Software.

83. Jadad AR, Cook DJ, Jones A, Klassen TP, Tugwell P, Moher M *et al.* Methodology and reports of systematic reviews and meta-analyses. *JAMA* 1998;**280**:278-80.

84. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomized controlled trials: the QUORUM statement. *Lancet* 1999;**354**:1896-900.

85. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D *et al.* Meta-analysis of observational studies in epidemiology. *JAMA* 2000;**283**:2008-12.

86. Jadad AR, Cook DJ, Browman GP. A guide to intrepreting discordant systematic review. *Can Med Assoc J* 1997;**156**:1411-6.

87. Egger M,.Smith GD. Bias in location and selection of studies. *BMJ* 1998;**316**:61-6.

88. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol* 1992;**45**:255-65.

89. Moher D, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P *et al.* Does quality of reports of randomized trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998;**352**:609-13.

90. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ *et al.* Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996;**17**:1-12.

91. Dinnes J, Deeks J, Sowden A, Song F, Altman D. A review of quality assessment tools for non-randomised intervention studies. *Int Soc Tech Assess Health Care* 2001; Philadelphia.

92. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M *et al.* The Newcastle-Ottawa Scale (NOS) for assess the quality of nonrandomised studies meta analyses. *3rd symposium for systematic reviews* 2000.

93. Reisch JS, Tyson JE, Mize SE. Aid to the evaluation of therapeutic studies. *Pediatrics* 1989;**84**:815-27.

94. Black NA,.Downs SH. The effectiveness of surgery for stress incontinence in women: a systematic review. *Br J Urol* 1996;**78**:497-510.

95. Zaza S, Wright-De Auguero LK, Briss PA, Truman BI, Hopkins DP, Hennessy MH *et al.* Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. Task Force on Community Preventive Services. *Am J Prev Med* 2000;**18**:44-74.

96. Taggart DP, D'Amico R, Altman DG. Effect of arterial revascularisation on survival: a systematic review of studies comparing bilateral and single internal mammary arteries. *Lancet* 2001;**358**:870-5.

97. Tugwell P, Wells GA, Peterson J, Welch V, Page J, Davison C *et al.* Do silicone breast implants cause rheumatologic disorders? *Arthr Rheum* 2001;**44**:2477-84.

98. Felson DT. Bias in meta-analytic research. *J Clin Epidemiol* 1992;**45**:885-92.

99. Lowe HJ,.Barnett GO. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA* 1994;**271**:1103-8.

100. Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. *Ann Intern Med* 1997;**127**:380-7.

101. Smith BJ, Darzins PJ, Quinn M, Heller RF. Modern methods of searching the medical literature. *Med J Aust* 1992;**157**:603-11.

102. Fleiss JL. Statistical methods for rates and proportions. 1981. Toronto Canada, John Wiley & Sons.

103. Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *New Engl J Med* 1992;**327**:248-54.

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

105. Jordan VC, Gapstur SM, Morton SC. Selective estrogen receptor modulation and reduction in risk of breast cancer, osteoporosis, and coronary heart disease. *J Natl Cancer Instit* 2001;**93**:1449-57.

106. Olsson H, Bladstrom A, Ingvar C, Moller TR. A population-based cohort study of HRT use and breast cancer in southern Sweden. *Br J Cancer* 2001;**85**:674-7.

107. Pike MC, Ross RK, Spicer DV. Problems involved in including women with simple hysterectomy in epidemiologic studies measuring the effects of hormone replacement therapy on breast cancer risk. *Am J Epidemiol* 1998;**147**:718-21.

108. Franceschi S. [Replacement therapy in menopause and risk for breast tumors]. [Italian]. *Annali dell Istituto Superiore di Sanita* 1997;**33**:207-11.

109. La Vecchia C, Negri E, Franceschi S, Favero A, Nanni O, Filiberti R *et al.* Hormone replacement treatment and breast cancer risk: a cooperative Italian study. *Br J Cancer* 1995;**72**:244-8.

110. Newcomb PA, Titus-Ernstoff KM, Egger M, Trentham-Dietz J, Baron J, Willett WC *et al.* Estrogen-progestin use in relation to risk of lobular and ductal breast cancer. *Society of Epidemiological Research Abstracts* 1998;**S64**:254.

111. Newcomb PA, Longnecker MP, Storer BE, Mittendorf R, Baron J, Clapp RW *et al.* Long-term hormone replacement therapy and risk of breast cancer in postmenopausal women. *Am J Epidemiol* 1995;**142**:788-95.

112. Ewertz M. Influence of non-contraceptive exogenous and endogenous sex hormones on breast cancer risk in Denmark. *Int J Cancer* 1988;**42**:832-8.

113. Ewertz M. Breast cancer risk in Denmark. *Acta Oncologica* 1993;**32**:595-615.

114. Li CI, Weiss NS, Stanford JL, Daling JR. Hormone replacement therapy in relation to risk of lobular and ductal breast carcinoma in middle-aged women. *Cancer* 2000;**88**:2570-7.

115. Stanford JL, Weiss NS, Daling JR, Daling J.R., Habel LA., Rossing MA. Combined estrogen and progestin hormone replacement therapy in relation to risk of breast cancer in middle-aged women. *JAMA* 1995;**274**:137-42.

116. Chen CL. A nested case-control study of hormone replacement therapy in relation to breast cancer. *Dissertation Abstracts* 1999;**60-08**:3885-179 p.

117. Kirsh VA. Hormone replacement therapy and the risk of breast cancer among post menopausal women. *Dissertation Abstracts* 1999.

118. Palmer JR, Rosenberg L, Clarke M, Miller DR, Shapiro S. Breast cancer risk after estrogen replacement therapy: results from the Toronto Breast Cancer Study. *Am J Epidemiol* 1991;**134**:1386-95.

119. Persson I, Thurfjell E, Bergstrom R, Holmberg L. Hormone replacement therapy and the risk of breast cancer. Nested case-control study in a cohort of Swedish women attending mammography screening. *Int J Cancer* 1997;**72**:758-61.

120. Kaufman DW, Palmer JR, de Mouzon J, Rosenberg L, Stolley PD, Warshauer ME *et al.* Estrogen replacement therapy and the risk of breast cancer: results from the case-control surveillance study. *Am J Epidemiol* 1991;**134**:1375-85.

121. Longnecker MP, Paganini-Hill A, Ross RK. Lifetime alcohol consumption and breast cancer risk among postmenopausal women in Los Angeles. *Cancer Epidemiology, Biomarkers & Prevention* 1995;**4**:721-5.

122. Hunt K, Vessey M, McPherson K, Coleman M. Long term surveillance of mortality and cancer incidence in women receiving hormone replacement therapy. *Br J Obstet Gynecol* 1987;**94**:620-35.

123. Archer DF, Bush T, Nachtigall LE. Re: effect of hormone replacement therapy on breast cancer risk: estrogen versus estrogen plus progestin. *J Natl Cancer Instit* 2000;**92**:1950-1.

124. Pike MC,.Ross RK. Re: effect of hormone replacement therapy on breast cancer risk: estrogen versus estrogen plus progestin. *J Natl Cancer Instit* 2000;**92**:1951-2.

125. Colditz GA,.Rosner B. Cumulative risk of breast cancer to age 70 years according to risk factor status: data from the Nurse's Health Study. *Am J Epidemiol* 2000;**152**:950-64.

126. Bergkvist L, Adami HO, Persson I, Hoover R, Schairer C. The risk of breast cancer after estrogen and estrogen-progestin replacement. *New Engl J Med* 1989; **321**:293-7.

127. Persson I, Yuen J., Bergkvist L, Schairer C. Cancer incidence and mortality in women receiving estrogen and estrogen-progestin replacement therapy-long term follow-up of a Swedish cohort. *Int J Cancer* 1996;**67**:327-32.

128. Schairer C, Byrne C, Keyl P, Brinton LA, Sturgeon S., Hoover R. Menopausal estrogen and estrogen-progestin replacement therapy and risk of breast cancer (United States). *Cancer Causes & Control* 1994;**5**:491-500.

129. Hennekens CH, Buring JE. Epidemiology in Medicine. Boston: Little, Brown and Company, 1987.

130. Nachtigall MJ, Smilen S, Nachtigall RD, Nachtigall RH, Nachtigall LE. Incidence of breast cancer in a 22-year study of women receiving estrogen-progestin replacement therapy. *Obstet Gynecol* 1992;**80**:827-30.

131. Aschengrau A, Paulu C, Ozonoff D. Tetrachloroethylene-contaminated drinking water and the risk of breast cancer. *Environ Health Perspec* 1998;**106**:947-53.

132. Magnusson C, Persson I, Adami HO. More about: effect of hormone replacement therapy on breast cancer risk: estrogen versus estrogen plus progestin. *J Natl Cancer Instit* 2000;**92**:1183-4.

133. Gordon M, Huang J. Osteoporosis. *Monograph Series on Age-related Diseases* 1995;**16**:1-30.

134. Yang CP, Daling JR, Band PR, Gallagher RP, White E, Weiss NS. Noncontraceptive hormone use and risk of breast cancer. *Cancer Causes & Control* 1992;**3**:475-9.

135. Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA* 1990;**263**:1385-9.

136. Sampson M, Barrowman NJ, Moher D, Klassen TP, Pham B, Platt R *et al.* Should systematic reviewers search Embase in addition to Medline? *JAMA* 2002; ***in press***.

137. Shapiro S. Meta-analysis/smeta-analysis. *Am J Epidemiol* 1994;**140**:771-8.

138. Willett WC, Colditz GA, Stampfer MJ. Postmenopausal estrogens--opposed, unopposed, or none of the above. *JAMA* 2000;**283**:534-5.

139. Col NF, Hirota LK, Orr RK, Eran JK, Wong JB, Lau J. Hormone replacement therapy after breast cancer: a systematic review and quantitative assessment of risk. *J Clin Epidemiol* 2001;**19**:2357-63.

140. Eden JA, Bush T, Nand S. A case-control study of combined continuous estrogen-progestin replacement therapy among women with a personal history of breast cancer. *Menopause* 1995;**2**:67-72.

141. Goodman MT, Cologne JB, Moriwaki H, Vaeth M, Mabuchi K. Risk factors for primary breast cancer in Japan: 8-year follow-up of atomic bomb survivors. *Prev Med* 1997;**26**:144-53.

142. Paganini-Hill A, Ross RK. Reliability of recall of drug usage and other health-related information. *Am J Epidemiol* 1982;**116**:114-22.

143. Daly E, Roche M, Barlow D, Gray A, McPherson K, Vessey M. Hormone replacement therapy: an analysis of benefits, risks and costs. *Br Med Bull* 1992;**307**:836-40.

144. Grodstein F. Can selection bias explain the cardiovascular benefits of estrogen replacement therapy? *Am J Epidemiol* 1996;**143**:979-82.

145. Willett WC. Diet and cancer: one view at the start of the millennium. *Cancer Epidemiology, Biomarkers & Prevention* 2001;**10**:3-8.

146. Lee IM, Rexrode KM, Cook NR, Hennekens C, Buring JE. Physical activity and breast cancer risk: the Women's Health Study (United States). *Cancer Causes & Control* 2001;**12**:137-45.

147. Smith-Warner SA, Spiegelman D, Yaun S-S. Alcohol and breast cancer in women: a pooled analysis of cohort studies. *JAMA* 1998;**279**:535-40.

148. Miller JN, Colditz GA, Mosteller F. How study design affects outcomes in comparisons of therapy. II: surgical. *Stats Med* 1989;**8**:466.

149. Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. *New Engl J Med* 1981;**309**:1358-61.

150. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New Engl J Med* 2000;**342**:1887-92.

151. Benson J, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *New Engl J Med* 2000;**342**:1878-86.

152. Linde K, Scholz M, Melchart D, Willich SN. Should systematic reviews include non-randomized and uncontrolled studies? The case of acupuncture for chronic headache. *J Clin Epidemiol* 2002;**55**:77-85.

153. Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Controlled Clinical Trials* 1990;**11**:339-52.

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

155. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999;**282**:1054-60.

156. Colditz GA. Relationship between estrogen levels, use of hormone replacement therapy, and breast cancer. *J Natl Cancer Instit* 1998;**90**:814-23.

157. Ferguson DJP, Anderson TJ. Morphological evaluation of cell turnover in relation to the menstrual cycle in the 'resting' human breast. *Br J Cancer* 1981;**44**:177-81.

158. Boyd NF, Lockwood GA, Byng JW. Mammographic densities and breast cancer risks. *Cancer Epidemiology, Biomarkers & Prevention* 1998;7:1133-44.

159. Cline JM, Soderqwist G, von Schoultz E. Effects of hormone replacement therapy on the mammary gland of surgically postmenopausal cynomolgus macaques. *Am J Obstet Gynecol* 1996; **174**:93-100.

160. Rutteman GR. Contraceptive steroids and the mammary gland: is there a hazard?--insights from animal studies. *Breast Cancer Research & Treatment* 1992;**23**:29-41.

161. Tavassoli FA. The influence of endogenous and exogenous reproductive hormones on the mammary glands with emphasis on experimental studies on rhesus monkeys. *Verh Dtsch Ges Pathol (abstract only)* 1997;**81**:514-20.

162. Hulka BS, Kaufman DG, Fowler WC Jr. Predominance of early endometrial cancers after long-term estrogen use. *JAMA* 1980;**244**:2422.

163. Oxman AD,.Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med* 1992;**116**:84.

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

165. Joffe MM, Byrne C, Colditz GA. Postmenopausal hormone use, screening and breast cancer: characterization and control of a bias. *Epidemiol* 2001;**12**:429-38.

166. Campagnoli C, Biglia N, Cantamessa C, Lesca L, Sismondi P. HRT and breast cancer risk: a clue for interpreting the available data. *Maturitas* 1999;**33**:185-90.

167. La Vecchia C, Negri E, Franceschi S, Parazzini F. Non-contraceptive oestrogens and breast cancer: an update. *Int J Cancer* 1992;**50**:161-2.

168. Pham B, Platt R, McAuley L, Sampson M, Klassen TP, Moher D. Detecting and minimizing publication bias: a systematic review of methods. *JAMA* 2002;in press.

169. Egger M, Zeilweger-Zahner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomized controlled trials published in English and German. *Lancet* 1997;**350**:326-9.

170. Moher D, Pham B, Klassen TP, Schulz KF, Berlin JA, Jadad AR *et al.* What contributions do languages other than English make on the results of meta-analyses? *J Clin Epidemiol* 2000;**53**:964-72.

171. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991;**337**:867-72.

172. McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses. *Lancet* 2000;**356**:1228-31.

173. Christiansen C, Christensen MS, Transbol I. Bone mass in postmenopausal women after withdrawal of oestrogen/gestagen replacement therapy. *Lancet* 1981; **329**:461.

174. Schneider ML, Barrett-Connor E, Morton DJ. Timing of postmenopausal estrogen for optimal bone mineral density. The Rancho Bernardo Study. *JAMA* 1997; 277:543-7.

175. Ventura P, Cagnacci A, Malmusi S, Panini R, Baldassari F, AranginoS *et al.* Continuous combined hormone replacement therapy with oral 17beta-estradiol and norethisterone acetate improves homocysteine metabolism in postmenopausal women. *Menopause* 2001;**8**:252-8.

176. Sarrel PM. Psychosexual effects of menopause: role of androgens. *Am J Obstet Gynecol* 1999;**180**:319-24.

177. Shoupe D. Androgens and bone: clinical implications for menopausal women. *Am J Obstet Gynecol* 1999;**80**:329-33.

178. Mosca L, Collins P, Herrington DM, Mendelsohn ME, Pasternak RC, Robertson RM *et al.* Hormone replacement therapy and cardiovascular disease. *Circulation* 2002;**104**:499-503.

179. Cummings SR, Eckert S, Krueger Ka, Grady D, Powles TJ, Cauley JA. The effect of raloxifene on risk of breast cancer in postmenopausal women: results from the MORE randomized trial. Multiple Outcomes of Raloxinfene Evaluation. *JAMA* 1999;**281**:2189-97.

180. Chen WY,.Colditz GA. Hormone replacement therapy, selective estrogen receptor modulators, and breast cancer risk. *Menopause* 1999;**6**:279-81.

181. Vickers MR. Women's International study of long duration oestrogen after menopause - WISDOM update. *J Br Meno Soc* 1999;**6**:159.

182. Million Women Collaborative Group. The million women study: design and characteristics of the study population. *Breast Cancer Research* 1999;**1**:73-80.

Appendix A-1 Newcastle Ottawa scale for case control studies

Study: Time: Rater's initials:

NEWCASTLE OTTAWA SCALE FOR CASE-CONTROL STUDIES	SCORE
SELECTION 1) Is the case definition adequate? a. Yes, with independent validation (eg medical/hospital records) ★ b. Yes, record linkage (eg ICD codes in database) or self report with no reference to primary record c. No description	
2) Representativeness of the cases a. Consecutive or obviously representative series of cases (eg. random sample)★ b. Potential for selection biases or not stated	
3) Selection of controls a. Community controls ★ b. Hospital controls c. No description	
4) Definition of controls a. No history of disease (endpoint)★ b. No description of source	
COMPARABILITY 1) Comparability of cases and controls on the basis of the design or analysis a. Study controls for AGE ★ b. Study controls for any additional factors ★ in analysis:	
OUTCOME 1) Ascertainment of exposure a. Secure record (eg. surgical records)★ b. Structured interview where blind to case/control status ★ c. Interview not blinded to case/control status— d. Written self report or medical record only e. No description	
2) Same method of ascertainment for cases and controls a. Yes ★ b. No	
3) Non-Response rate a. Same rate for both groups ★-- b. Non respondents described c. Rate different and no designation	

Summary score

SELECTION (4★)	COMPARABILITY (2★)	EXPOSURE (3★)

Appendix A-2. Newcastle Ottawa scale for cohort studies

Study: Time: Rater's initials:

NEWCASTLE OTTAWA SCALE FOR COHORT STUDIES	SCORE
SELECTION 1) Representativeness of the exposed cohort a. Truly representative of the average postmenopausal woman in the community ★ b. Somewhat representative of the average postmenopausal woman in the community ★ c. Selected group of users (nurses, volunteers) d. No description of the derivation of the cohort	
2) Selection of the non exposed cohort a. Drawn from the same community as the exposed cohort ★ b. Drawn from a different source c. No description of the derivation of the exposed cohort	
3) Ascertainment of exposure a. Secure records (doctor's record)★ b. Structured interview ★ c. Written self report d. No description	
4) Demonstration that outcome of interest was not present at start of study a. Yes ★ b. No	
COMPARABILITY 1) Comparability of cohorts on the basis of the design or analysis a. Study controls for AGE ★ b. Study controls for any additional factor ★	
OUTCOME 1) Assessment of outcome a. Independent blind assessment ★ b. Record linkage ★ c. Self report d. No description	
2) Was follow-up long enough for outcomes to occur (~ 10 years) a. Yes ★ b. No	
3) Adequacy of follow up of cohorts a. Complete follow up – all subjects accounted for ★ b. Subjects lost to follow up unlikely to introduce bias—small number lost (< 20%) or description provided of those lost ★ c. Follow up rate > 80% and no description of those lost d. No statement	

Summary Score

SELECTION (4★)	COMPARABILITY (2★)	EXPOSURE (3★)

Appendix B. An evaluation of the Newcastle Ottawa Scale: ***A scale designed for the quality assessment of nonrandomized studies used in meta-analysis. by*** **LM Brodsky**

Abstract

Objectives: To evaluate the validity and reliability of the Newcastle Ottawa Scale (NOS)—a scale developed for assessing the quality of nonrandomized studies.

Methods: Face validity of the NOS was examined by reviewing the two forms of the scale and checking the wording of each of the items and determining if the answers match the question stem.

Criterion validity was examined by comparing NOS with an established scale (Black and Downs). The quality of twenty studies (10 case controls; 10 cohorts) were assessed using both scales and the level of association between the scores was determined by computing intra-class correlations.

Inter-rater reliability was determined using two raters that assessed 10 cohort studies and 10 case-control studies. Intra-class correlation coefficients were calculated to measure the level of agreement.

Evaluator burden was also measured by recording the time required to evaluate a study using the two different scales. Time to score was determined for ten case control and ten cohort studies.

Results: NOS demonstrated very good face validity. The scale was clearly written and was easy to follow and all of the responses matched the question stem. There were some ambiguities in the phrasing of the answers, but these could be readily modified.

The NOS scoring compared favourably with the scoring of the more established Black and Downs' scale. While there was an excellent level of agreement with the cohort studies (ICC = 0.88), the level of agreement with case-control studies was weaker (ICC= 0.62).

The NOS demonstrated a high degree of reliability as measured by inter-rater reliability. Inter-rater reliability was high for both case-control studies (ICC = 0.82) and cohort studies (ICC = 0.94).

The NOS was easier and faster to use than the Black and Downs' scale. The average time to evaluate a case control study using the NOS ($x = 7.46 \pm 0.96$ minutes) was significantly less than the time it took using the Black and Downs' scale ($x = 16.7 \pm 1.7$ minutes; $t = 6.28$, $df = 9$; $p < 0.001$). Similarly, the average time to score a cohort study using the NOS ($x = 8.22 \pm 1.14$ minutes) was significantly less than the average scoring time using the Black and Downs' scale ($x = 15.72 \pm 3.14$; $t = 8.42$; $df = 9$; $p < 0.001$).

Conclusions: The NOS was a very good scale for assessing the quality of nonrandomized studies. It demonstrated strong face and criterion validity. Its reliability was excellent and the evaluator burden was low compared to the more established Black and Downs' scale. With a few minor changes, it will provide meta-analysts with a valuable tool for assessing the methodological quality of nonrandomized studies.

Healthcare professionals, researchers, consumers and policy makers are inundated with increasing volumes of medical reports. To effectively manage all of this information, individuals often rely on the results from systematic reviews of the literature. A systematic review refers to any type of review that has been prepared using strategies to avoid bias (Egger and Smith 1997). In the early 1990's, a group of clinicians, epidemiologists and other health professionals formed the Cochrane Collaboration, an organization whose goal is to prepare, maintain and disseminate systematic reviews on the effects of health care. A Cochrane Systematic Review consists of seven well-defined steps which include a structured comprehensive search of the literature for all studies that pertain to the question of interest (eg. Does hormone replacement therapy (HRT) reduce the risk of heart disease among women?). A critical appraisal and synthesis of the data extracted from each study is followed by a statistical analysis that combines the results of the studies (Mulrow and Oxman 1997). This statistical pooling of the data extracted during the systematic review is termed a meta-analysis (Last 2001). The resulting quantitative summary statistic provides a consensus on the overall results.

There are several ways that a meta-analysis can be used. For instance, it can estimate the degree of benefit associated with a particular study treatment; assess the amount of variability between studies; identify study characteristics that are associated with effective treatment effects and increase the power to detect an overall treatment effect (Normand 1999). Because combining several studies increases the sample size, a meta-analysis can also strengthen the statistical power necessary to observe small but clinically important effects (Friedman et al 1998). Therefore, the results from a meta-analysis can

be used as important source of information for healthcare providers trying to make decisions about their patients, practices, hospitals and health care systems (Sackett et al 1991).

Despite its widespread use, meta-analysis continues to be a controversial technique. For instance, Jadad et al 1997 report that meta-analyses addressing the same therapeutic question often reach opposite conclusions. Such discordance will present difficulties for the decision-makers who rely on these reviews to help them make choices among alternative health care interventions. Egger and Smith 1998 argue that these discrepancies are due to various types of selection bias inherent in the studies chosen for inclusion in the meta-analysis. Detsky et al 1992 propose that variation in the quality of the trials that are included in the meta-analysis can also lead to bias. They report that many researchers are uneasy about combining the results from studies of differing qualities. Since poor quality studies are more likely to have errors in estimating the outcomes of interest, incorporating them into an meta-analysis could lead to Type I or Type II error.

Even more controversial is the application of meta-analytic techniques to nonrandomized (observational) studies. The main problem with nonrandomized studies involves the composition of the study groups. In an RCT, the clinicians can randomly and blindly allocate participants into treatment and control groups. When designing nonrandomized studies, the participants are assigned to their groups based on their prior "experiences". In cohort studies, they are either exposed or unexposed to the intervention. In case/control studies, they already have the outcome (case) or not

(control). This contributes to much of the bias associated with non randomized studies and brings into question the rigour of their findings.

Shapiro 1994 opposes the use of meta-analysis for nonrandomized studies. He argues that if the same systematic biases are present in a number of studies, then the meta-analysis will reinforce the biases and lead to "spurious statistical stability". While many believe there is still merit in conducting meta-analyses on nonrandomized studies (eg. Greenland 1994; Petitti 1994), the need to assess the quality of the studies included in the meta-analysis is critical.

If meta analyses are to be used as a clinical decision making tool, it is essential that they follow guidelines to minimize bias. One way to deal with bias is to assess the quality of studies before their inclusion in a meta-analysis. For instance, several studies have evaluated the role that quality assessment plays in meta-analysis results (Clemens et al 1983; Moher et al 1998). In both studies, they found that inclusion of poor quality studies in their analysis altered the results considerably. By assessing the quality of a study, the researcher is provided with an indication that the results are a valid estimate of the truth and thus should be incorporated into a meta-analysis (Jadad et al 1996).

Numerous scales and checklists have been developed to assist with the quality assessment of studies. However most of these scales have been developed to assess the quality of randomized controlled trials (RCTs). For instance, as of 1994, 25 scales and 9 checklists have been developed to assess RCT quality. While checklists are useful for the author to determine if their report contains the necessary information, the scale can provide a quantitative index of the study quality. A high quality study would imply that its reported methodology and results are free of bias (Moher et al 1995).

Several authors claim to have developed scales to assess the quality of nonrandomized studies (eg. Zola et al 1989; Steibe et al 1990; Bass et al 1993; Hadorn et al 1996; Heneghan et al 1996). Because these scales have been designed primarily to evaluate RCTs, their assessment of nonrandomized studies is compromised. In response to these shortcomings, Wells et al (2000) developed the NOS. It consists of two difference scales, one that can be used to assess the quality of cohort studies and the other, case-control studies.

The NOS evaluates a study using the following three perspectives: Selection of the study groups; Comparability of the two groups of individuals (exposed and non-exposed in the cohort studies; cases and controls in the case-control studies); and Ascertainment of outcome (cohort studies) or exposure (case-control studies).

The scale is further organized into numbered items. While the Comparability consists of only one item, the Selection and Ascertainment perspectives contain four and three items respectively. These eight items enable the reader to assess how well the study was performed. The scoring consists of a star system; stars are awarded if the study meets a particular criteria featured in the item; the more stars awarded then the higher the study quality. For instance, if the Ascertainment of exposure was obtained through medical records, a star is awarded but if it was obtained from a written self report, no star is given. A study can gain a maximum of four stars for Selection characteristics; two stars for Comparability and three stars for Ascertainment of exposure/outcome. These scores are summed within each of the three perspectives, but a total score is not computed.

To date, this scale has been used in several meta-analyses (association of breast cancer with HRT; association of connective tissue disease with silicone breast implants). While the scale has been refined as a result of these projects, no formal evaluation of NOS has taken place. The aim of this study is to evaluate the validity and reliability of this scale and determine what role it could play in assessing the quality of nonrandomized studies.

Methods

Face validity—The items in the two versions of the NOS were carefully reviewed to determine if the items appear on the surface to be assessing what they set out to measure. The items were examined to ensure that the phrasing of each of the items was unambiguous. The answers provided for each of the items were also assessed relative to the question stem to verify that they matched.

Criterion validity—In order to determine if the NOS produces results which correspond with those obtained from a well established scale, the NOS was compared to one developed by Black and Downs 1997. The quality of 10 case control studies was assessed using the Black and Downs' scale and the NOS version for case controls studies. Scoring results for each study using the two scales was recorded on separate forms designed specifically for this purpose (see Appendix 2). To reduce bias, all of the studies were first scored using the NOS and then a week later scored using the Black and Downs' scale. This prevented the reviewer from remembering the relative scores of the studies and biasing their scoring on the second scale. For example if a study received a high

NOS score, then remembering it as a high quality study may have biased the scoring towards the upper end of the Black and Downs' scale. The same procedure was then repeated for 10 cohort studies, except that the NOS designed specifically for cohort studies was used.

To determine if the NOS scores corresponded with the Black and Downs scores, the total stars/points from each scale were tallied. The level of association was determined using an interclass correlation coefficient (SPSS version 9.0). Two separate ICCs were calculated, one involving the comparison of the 10 case control studies and the other for the 10 cohort studies.

Inter-rater reliability—Two raters (LMB, Joan Peterson) independently evaluated the same 20 articles: 10 case control studies and 10 cohort studies using the appropriate version of NOS. They recorded their responses on separate forms (Appendix 2). After the completion of the rating process, the two sets of scores were compared. Two separate ICCs were calculated (one for case-control studies; one for cohort studies) to determine the level of agreement between the rater's scoring.

Measure of burden— During the assessment of criterion validity, the time required to score each of the 20 different studies with the NOS and Black and Downs scale was measured using a stopwatch. The mean scoring time was determined for the 10 case control studies and 10 cohort studies for each of the two scales and compared using paired t tests.

Results

Face validity—The two versions of the NOS were examined to determine if they are relevant to the quality assessment of nonrandomized studies, and to assess if the presentation of the items are unambiguous. One of the major problems with nonrandomized studies is the bias often associated with the design of the study. A biased study would indicate poor methodological quality and its results would have little meaning. The goal of the NOS is to assess the quality of nonrandomized studies by evaluating how well these biases are controlled.

Each version of the NOS is divided into three criteria: Selection; Comparability and Outcome/exposure (Appendix 2). The Selection perspective in an essential component for nonrandomized studies. The participants in nonrandomized studies are not selected randomly (as the name suggests) and thus there is the possibility that there could be some bias introduced into the study based on how participants were selected. It is essential to control for this bias otherwise the results of the study may be questionable. If the participants in the study are a select group, then it will be difficult to broadly apply the conclusions of the study to other populations and this would compromise the quality of the study. The NOS distinguishes between studies with participants that are "representative" of the community (close to a random sample) and studies where the participants are part of a select group.

The Comparability perspective is also essential for assessing the bias often present in the design or analysis of studies. In nonrandomized studies, individuals are assigned to groups based on their exposure (exposed/not exposed) or outcome (case—outcome/control—no outcome). This could contribute to bias because the exposed

groups could have a different set of circumstances or experiences than the non-exposed group and you would be unable to discern if the differences in the results of the study were due to this or the exposure.

The Comparability perspective attempts to assess this by distinguishing between studies that control for the main risk factors (in either design or analysis). The scale also offers a degree of flexibility in the assessment because it permits the evaluator to identify the most important risk factors and assess them accordingly. This broadens the applicability of the NOS to any meta-analysis of nonrandomized studies.

The Exposure/Outcome perspective focuses on how accurately a particular study assesses the exposure or outcome. Since these characteristics determine the classification of participants, it is essential that the assessment is accurate. The NOS distinguishes between cohort studies that use independent assessment of outcome and those that rely on self-report. They also distinguish between case-control studies that use a secure exposure record (medical records) versus self-report. The NOS deals with these biases by ensuring that those studies that rely on self-report receive a lower quality score than those that rely on secure records.

The scale was clearly written and easy to follow and all of the responses matched the question stem. A few of the items were difficult to interpret and these would benefit from further clarification. The main problem involves the subjective nature of several of the questions. For instance, question 2 of the case-control version asks if the cases are "representative" and question 1 on the cohort version asks if the cohort is "truly or somewhat representative". This terminology is subjective and should be modified to avoid misinterpretation. The last question in the cohort version asks if the follow up was

"adequate". This needs to be more clearly defined because it is subject to interpretation. Another problem concerns exposure assessment, a key components in cohort and case control studies. A star is awarded if the ascertainment was obtained through a structured interview but not a written self-report. Evaluators may interpret a questionnaire that the respondent fills out on their own as a self-report while others may interpret it as a structured interview. This requires further clarification.

Criterion validity—Figures 1 and 2 summarize the results of the criterion validity assessment. Since the NOS ranges from 0-9 and the Black and Downs' scale (B&D) ranges from 0-27, the scores were standardized for the illustrative purposes.
With the cohort studies, the trend indicates that studies judged as low quality with the NOS also scored low on the Black and Downs' scale, and the similar trend was noted for the high quality studies. This trend was not as evident with the case-control studies.

There was some difficulty encountered in scoring case-control studies using the Black and Downs scale. Six of the 27 questions in this scale could not be applied to a case-control study. For instance, it was not possible to assess the loss to follow up (2 questions), the length of follow up, compliance with intervention, and random and concealed allocation (questions 9,26, 17, 19, 23, 24; see appendix).

It was easier to score a cohort study using the Black and Downs' scale, but there were still questions that were difficult to answer. Compliance with intervention and random/concealed allocation also presented problems for cohort studies. However, since the follow up could be assessed, more questions could be answered overall. This difference was reflected in the validity scores.

The level of association between the quality assessment scores determined by the two scales was estimated using intra-class correlation coefficients. The ICC for case control studies was 0.61 and for cohort studies, 0.88.

Inter-rater reliability—Figures 3 and 4 illustrate a comparison of the scoring results obtained by the two raters. The two raters were quite similar in scoring. They awarded the identical NOS score for 9 of the 20 studies evaluated. Of the remaining 11 studies, the score differed by only one point for 9 of the studies and by two points for 2 of the studies.

Tables 1 and 2 display the two raters' scoring for the case control studies and tables 3 and 4 display the two raters' scoring for the cohort studies for each of the nine questions of the NOS. It is interesting to note that when the two raters awarded the same score to a study, the scoring was identical for each of the nine questions. For instance, Rater 1 awarded the Guevara 1992 paper a total score of six points; three points for Selection, two points for comparability and one point for exposure. Rater 2 also awarded a total of six points to this study, and the identical score for each of the nine questions. There was no instance of a study being awarded the same total score with a different combination of points. For instance one rater could have awarded 2 points for each of the three categories for a total of 6 points while the other rater could have awarded 4 points for selection; 0 points for comparability and 2 points for exposure for an identical total of 6. This would be an instance of same total score but different scoring pattern. Thus the total score and scoring pattern was consistent for all of the twenty studies

The inter-class correlation coefficient for the case control studies was 0.84. The ICC for the cohort studies was 094. This indicates that there is very good inter-rater reliability with the NOS.

Measure of burden—Quality assessment using the NOS took less time than the Black and Downs' scale for all 20 studies evaluated (Figures 5 and 6). The average time to evaluate a case control study using the NOS ($x = 7.46 \pm 0.96$ minutes) was significantly less than the time it took using the Black and Downs' scale ($x = 16.7 \pm 1.7$ minutes; $t = 6.28$, $df = 9$; $p < 0.001$). Similarly, the average time to score a cohort study using the NOS ($x = 8.22 \pm 1.14$ minutes) was significantly less than the average scoring time using the Black and Downs' scale ($x = 15.72 \pm 3.14$; $t = 8.42$; $df = 9$; $p < 0.001$). Therefore the measure of burden is considerably less when the evaluator uses the NOS. It is important to note that actual scoring time for the Black and Downs' scale was underestimated for all 20 studies since the assessment of the power component (question 27) was omitted.

Discussion

Face validity--While face validity is usually evaluated by experts (Streiner and Norman 1995), it could also be useful to obtain the perspective of a person with no previous experience with the scale. They would be free of preconceived notions of the relationship between "what is being said by the scale" and "what is really meant". This should result in an assessment free of bias.

The fact that the NOS distinguishes between studies with participants that are "representative" of the community (close to a random sample) and studies where the

participants are part of a select group implies that the bias associated with a nonrandom sample has been addressed. A study that controls for this potential selection bias is viewed to have higher methodological quality compared to one that fails to do so.

In nonrandomized studies, there can be a selection bias associated with the individuals assigned to the two groups. In RCTs, individuals are randomly assigned to the different arms of the trial (eg. treatment, control groups). There is usually no difference between the individuals in the separate groups as a result of this randomization and thus any differences between outcomes are due to the differences in treatment that were received. However, in nonrandomized studies, there could be a bias in a case/control study, such that the cases are older on average than the controls. This can present a problem in the interpretation of the results and compromise the findings of the study. The NOS deals with this problem by distinguishing between studies that control for the main risk factors (in either design or analysis) and those that fail to do so. The fact that studies that do control for the main risk factors receive a higher score (by up to two points) than those that do not demonstrates that the NOS addresses this potential bias.

Two of the main biases associated with nonrandomized studies are recall bias (case control studies) and classification bias (cohort studies). If one relies on self-reporting, there is always the issue of how well the individual remembers. The fact that the NOS deals with these biases by ensuring that those studies that rely on self-report receive a lower quality score than those that rely on secure records demonstrates that the control of bias is reflected in the score.

The results of the face validity assessment revealed that the NOS permits the evaluator to assess the quality of the case control and cohort studies by how well the

study controls the biases inherent in design and analysis. For each of the scale's three perspectives, studies that controlled the bias through the careful selection of the participants, the controlling of risk factors and the ability to verify the records, receive a higher score than those studies that ignored these issues. Overall, the NOS items were found to be clear, concise and easy to understand. There were only a few ambiguities in the wording of the NOS items that could be easily rectified. Based on these findings, the NOS was determined to have excellent face validity.

Criterion Validity—The quality assessment scores of studies using the NOS was very similar to the more established Black and Downs' scale. The ICC scores suggest that these scoring systems are comparable, especially for the cohort studies. Based on the results of this study, the NOS was found to be superior to the Black and Downs' scale for evaluating nonrandomized studies. Like many of the other scales that have been developed for this purpose, the questions on Black and Downs' scale are more applicable to scoring RCTs than nonrandomized studies. In particular, the Black and Downs' scale fared particularly poorly when evaluating case-control studies. There were at least seven questions that were not applicable to case control studies and at least four that could not be applied to cohort studies. Therefore, it appears that the Black and Downs' scale is not a very valid tool for assessing quality of case-control studies While it could be modified for cohort studies, significant changes would be necessary in order to permit proper evaluation of all nonrandomized designs.

Inter-rater reliability—The NOS scores of the two raters were very similar, indicating a very strong inter-rater reliability. This is an important consideration for several reasons. First of all, most meta-analytic guidelines suggest that two raters assess the quality of

each study. The fact that the scores were so similar not only makes the NOS a valuable tool for evaluation within studies, it would also serve to enhance the comparability of quality assessment between different meta-analyses. One area that meta-analysis could benefit from is greater standardization of methods. If the same tool was used by different meta-analysts to assess different meta-analyses, there would be greater consistency in the quality of meta-analytic research.

Measure of burden—It took significantly less time to score a study with the NOS than with the Black and Downs' scale. This is likely due to the fact that the NOS has fewer questions and is easier to understand. One of the reasons that new scales are developed is to provide an instrument that is more practical, more economical or produces results faster (Streiner and Norman 1995). Therefore, the NOS will be a useful tool for assessing the quality of nonrandomized studies, particularly when the meta-analysis involves a large number of both case-control and cohort studies to evaluate.

Conclusions-- The results from this study reveal that the NOS would be an excellent tool for evaluating the methodological quality of nonrandomized studies. It has strong face validity, effectively addressing the major biases inherent in nonrandomized studies. In addition, it is easy to understand, thus simplifying the scoring process. The fact that the NOS has two versions, each one designed specifically for the evaluation of case-control or cohort studies, allows for a more accurate depiction of the study quality of nonrandomized design than the established Black and Downs' scale. In addition, the NOS was found to have high reliability. This is important because one of the criteria for

meta-analyses requires two raters for study evaluation. The fact that there is good inter-rater agreement makes the NOS a good tool for meta-analytic research. Finally, the evaluator burden is significantly less. This is an important consideration because meta-analyses can involve reviewing upwards of fifty articles. The NOS is a scale that is clear, reliable, valid and efficient to use. It will be a most welcome tool for meta-analytic research.

Literature cited

Barrett-Connor E, Friedlander NJ. 1993. Dietary fat, calories and the risk of breast cancer in postmenopausal women: a prospective population based study. J Am Coll Nutrit 12 (4): 390-399.

Bass JL, Christoffel KK, Widome M, Boyle W, Scheidt P, Stanwick R, and Roberts K 1993. Childhood injury prevention counseling in primary care settings: a critical review of the literature. Pediatrics *92*, 544-550.

Black NA and Downs SH 1996. The effectiveness of surgery for stress incontinence in women: a systematic review. Brit J Urol *78*, 497-510.

Clemens JD, Chuong JH, and Feinstein AR 1983. The BCG controversy: a methodological and statistical reappraisal. JAMA *249* , 2362-2369.

Colditz GA, Hankinson SE, Hunter DJ, Willett WC, Manson JE, Stampfer MJ, Hennekens C, Rosner B, Speizer FE. 1995. The use of estrogens and progestins and the risk of breast cancer in postmenopausal women. New England Journal of Medicine *332*, 1589-1593.

Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, and L'Abbe KA 1992. Incorporating variations in the quality of individual randomized trials into meta-analysis. J Clin Epidemiol *45*, 255-265.

Egger M and Smith GD 1997. Meta-analysis: potential and promise. BMJ *315*, 1371-1374.

Egger M and Smith GD 1998. Bias in location and selection of studies. BMJ *316*, 61-66.

Ewertz M. 1993. Breast cancer in Denmark. Acta Oncologica 32(6): 595-615.

Folsom AR, Mink PJ, Sellers TA, Ching-Ping J, Zyeng W, Potter JD 1995. Hormonal replacement therapy and morbidity and mortality in a prospective study of postmenopausal women. Am J Publ Health 85(8): 1128-1132.

Friedman LM, Furberg CD, and DeMets DL 2001. Fundamentals of Clinical Trials. (New York: Springer-Verlag).

Gambrell RD 1986. Cancer and the use of estrogens. Int J Fertility 31(2): 112-122.

Greenland S 1994. Can Meta-analysis be salvaged? Am J Epidemiol, 140: 783-787.

Guevara EB. 1991 ERT and breast cancer. Dissertation Abstracts International 52(12): 6338B.

Hadorn DC, Baker D, Hodges JS, Hicks N 1996. Rating the quality of evidence for clinical practice guidelines. J Clin Epidemiol 49, 749-754.

Harris RE, Namboordiri KK, Wynder EL. 1992. Breast cancer risk: effects of ERT and body mass. J Natl Cancer Inst 84(20): 1575-1582.

Heneghan A, Horowitz SM, and Leventhal JM 1996. Evaluating intensive family preservation programs: a methodological review. Pediatrics 97: 535-542.

Hennekens CH and Buring JE 1987. Epidemiology in Medicine., Mayrent SL, ed. (Boston: Little, Brown and Company).

Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ, and McQuay HJ 1996. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Controlled Clinical Trials 17: 1-12.

Jadad AR, Cook DJ, and Browman GP 1997. A guide to intrepreting discordant systematic review. Can Med Assoc J 156: 1411-1416.

Last JM 2001. A dictionary of epidemiology. (New York: Oxford University Press).

Lauritzen C, Meier F. 1984. Risks of endometrial and mammary cancer morbidity and mortality in long-term oestrogen treatment. *In* The Climateric: an update. Edited by H. Van Herendael. MTP Publishers. London.

Mills PK, Beeson WL, Phillips RL, Fraser GE. 1989. Prospective study of exogenous hormone use and breast cancer in Seventh-day Adventists. Cancer 64:591-597.

Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, and Walsh S 1995. Assessing the quality of Randomized controlled trials: an annotated bibliography of scales and checklists. Controlled Clinical Trials 16: 62-73.

Moher D, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, and Klassen TP 1998. Does quality of reports of randomized trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 352: 609-613.

Mulrow CD and Oxman AD 1997. Cochrane Collaboration Handbook. (Oxford: The Cochrane Collaboration).

Nachtigall MJ, Smilen SW, Nachtigall RD, Nachtigall RH, Nachtigall LE. 1992. Incidence of breast cancer in a 22 year study of women receiving estrogen-progestin replacement therapy. Obstet Gynecol 80: 827-30.

Newcomb PA, Longnecker MP, Storer BE, Mittendorf R, Baron J, Clapp RW, Bogdan G, Willett WC. 1995. Long-term HRT and risk of breast cancer in postmenopausal women. Am J Epidemiol 142 788-795.

Normand SL 1999. Tutorial in Biostatistics. Meta-analysis: formulating, evaluating, combining and reporting. Statist Med 18:321-359.

Palmer JR, Rosenberg L, Clarke EA, Miller DR, Shapiro S. 1991 Breast cancer risk after ERT: results from the Toronto Breast Cancer Study. Am J Epidemiol 134(12): 1386-1395.

Petitti DB 1994. Of babies and bathwater. Am J Epidem 140: 779-782.

Sackett DL, Haynes RB, Guyatt GH, and Tugwell P 1991. How to read reviews and economic analyses. In Clinical Epidemiology: A basic science for clinical medicine, (Boston: Little, Brown and Company), pp. 379-391.

Shapiro S 1994. Meta-analysis/smeta-analysis. Am J Epidemiol 140: 771-778.

Stanford JL, Weiss NS, Voigt LF, Daling JR, Habel LA, and Rossing MA 1995. Combined estrogen and progestin HRT in relation to risk of breast cancer in middle-aged women. JAMA 274: 137-142.

Streiner DL, Norman GR. 1995. Health measurement scales: a practical guide to their development and use. 2nd edition. Oxford University Press. New York.

Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, and Tugwell P (2000). The Newcastle-Ottawa Scale (NOS) for assess the quality of nonrandomised studies meta analyses. 3rd symposium for systematic reviews.

Zola P, Volpe T, Castelli G, Sismondi P, Nicolucci A, Parazzine F, and Liberati A (1989). Is the published literature a reliaable guide for deciding between alternative treatments for patients with early cervical cancer? Int J Rad Oncol Biol Phys *16*, 785-797.

Table 1. Inter-rater reliability: results from the quality assessment of case-control studies as scored by Rater 1

NOS criteria:	**Selection**				**Comparability**		**Exposure**			
Study, Year	Case definition Adequate	Cases representative	Community controls	Controls: No history of disease	Study controls for age	Study controls for another risk	Secure record of EPT history	Same for cases and control	Same non-response rate	Total score /9
Ewertz 1988	★	★	★	★	★	★	★	★	★	9★
Kaufman 1991	★	★	--	★	★	★	★	★	★	8★
Yang 1992	★	★	★	★	★	★		★	★	8★
Newcomb 1995	★	--	★	★	★	★	★	★	★	8★
Stanford 1995	★	★	★	★	★	★	--	★	★	8★
La Vecchia 1995	★	★		★	★	★	★	★	★	8★
Boice 1995	★	★	★	★	★	★		★	--	7★
Harris 1992	★		★		★	★		★		5★
Guevara 1992	★	★	★		★	★		-★-		6★

Table 2. Inter-rater reliability: results from the quality assessment of case-control studies as scored by Rater 2

NOS criteria:	**Selection**				**Comparability**		**Exposure**			
Study, year	Case definition Adequate	Cases representative	Community controls	Controls: No history of disease	Study controls for age	Study controls for another risk	Secure record of EPT history	Same for cases and control	Same non-response rate	Total score /9
Ewertz 1988	★	★	★	★	★	★		★		7★
Kaufman 1991		★		★	★	★		★	★	7★
Yang 1992	★		★	★	★	★		★	★	7★
Newcomb 1995	★		★	★	★	★	★	★	★	8★
Stanford 1995	★	★	★	★	★	★		★		7★
La Vecchia 1995	★			★	★	★		★	★	6★
Boice 1995	★	★	★	★	★	★		★		7★
Harris 1992	★		★		★	★		★		5★
Guevara 1992	★	★	★		★	★		★		6★

Table 3. Inter-rater reliability: results from the quality assessment of cohort studies as scored by Rater 1

NOS criteria:	**Selection**				**Comparability**		**Outcome**			
Study, year	Exposed group representative of community	Selection of the non exposed	Secure EPT exposure records	Outcome not present at start	Study controls for age	Study controls for another risk	Assessment of outcome is blind	Adequate follow-up for outcome	Adequate follow-up of cohorts	Total score /9
Gambrell 1986		★	★				★			3★
Mills 1989		★	★						★	3★
Barrett-Connor 1993		★	★	★	★	★	★	★	★	8★
Lafferty 1994		★	★	★	★		★		★	6★
Risch 1994	★	★	★	★			★	★	★	7★
Schairer 1994	★	★	★	★	★	★	★		★	8★
Colditz 1995		★	★	★	★	★	★	★		7★
Folsom 1995	★	★	★	★	★	★	★		★	8★
Lauritzen 1984		★	★		★	★				4★
Nachtigall 1992		★	★	★	★			★	★	6★

Table 4. Inter-rater reliability: results from the quality assessment of cohort studies as scored by Rater 2.

NOS criteria:	**Selection**				**Comparability**		**Outcome**			
Study, year	Exposed group representative of community	Selection of the non exposed	Secure EPT exposure records	Outcome not present at start	Study controls for age	Study controls for another risk	Assessment of outcome is blind	Adequate follow-up for outcome	Adequate follow-up of cohorts	Total score /9
Gambrell 1986		★	★				★	★		4★
Mills 1989		★							★	2★
Barrett-Connor 1993		★	★	★	★	★	★	★	★	8★
Lafferty 1994		★	★	★	★				★	5★
Risch 1994	★	★	★	★			★	★	★	7★
Schairer 1994		★	★	★	★	★	★		★	7★
Colditz 1995		★	★		★	★	★	★	★	7★
Folsom 1995	★	★		★	★	★	★		★	7★
Lauritzen 1984		★	★		★	★		★		5★
Nachtigall 1992		★	★	★	★		★	★	★	7★

Figure 1. A comparison of NOS and B&D scores for Case control studies

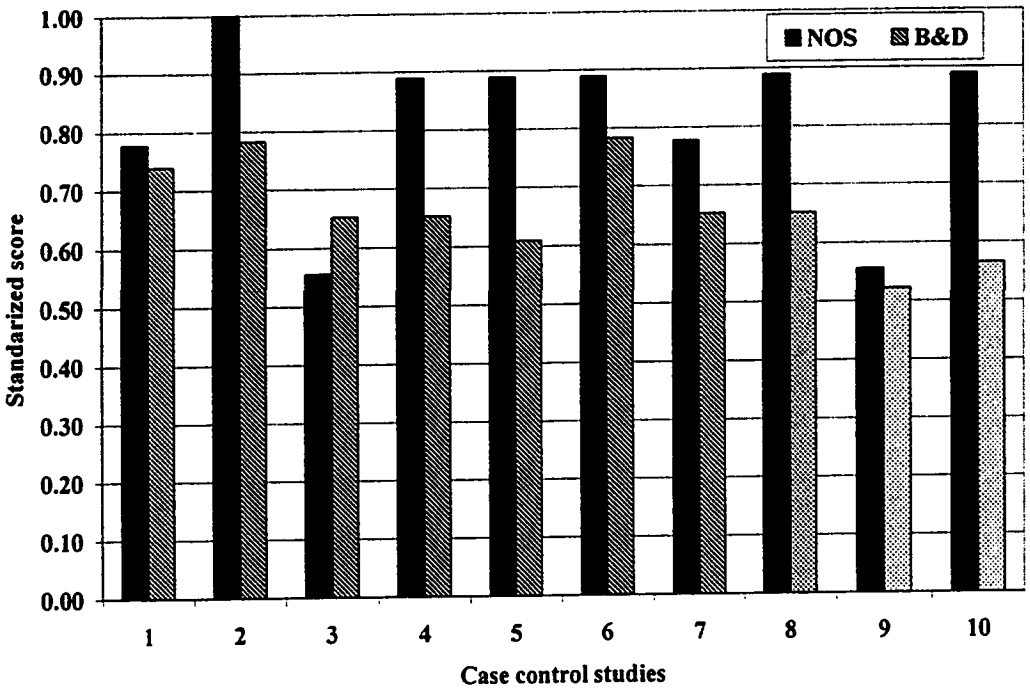


Figure 2. A comparison of NOS and B&D quality scores for cohort studies

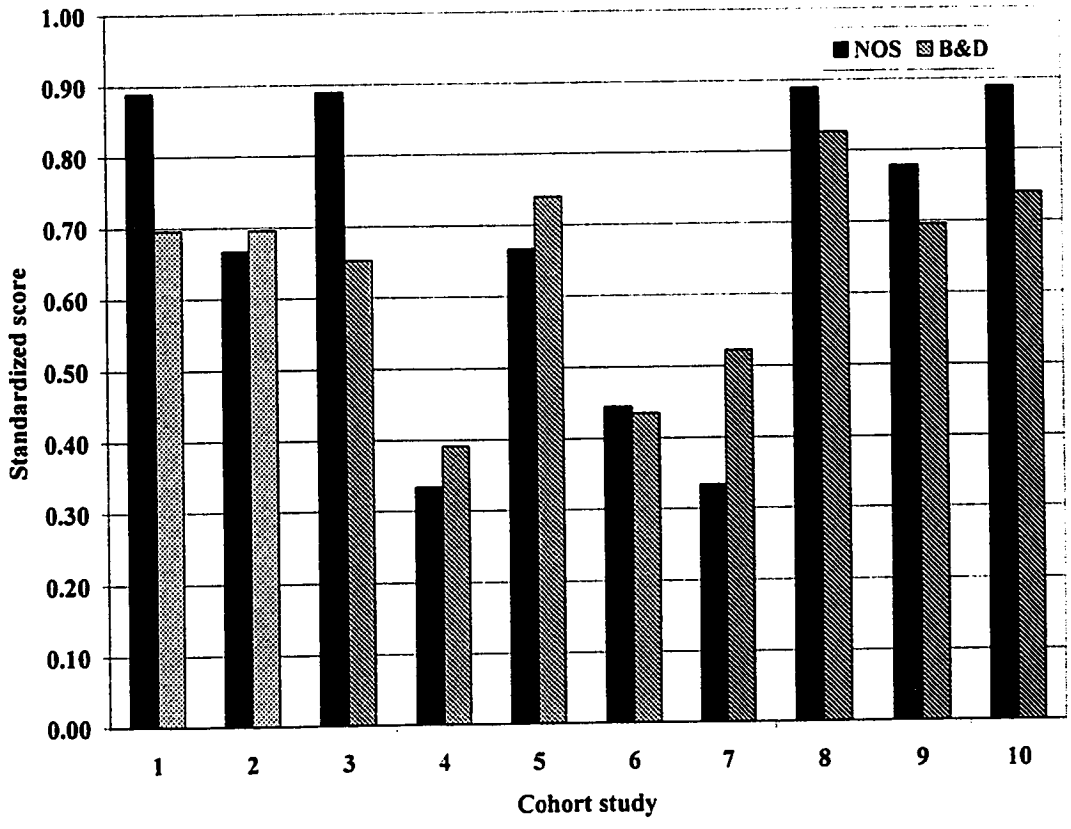


Figure 3. Inter-rater reliability for case control studies

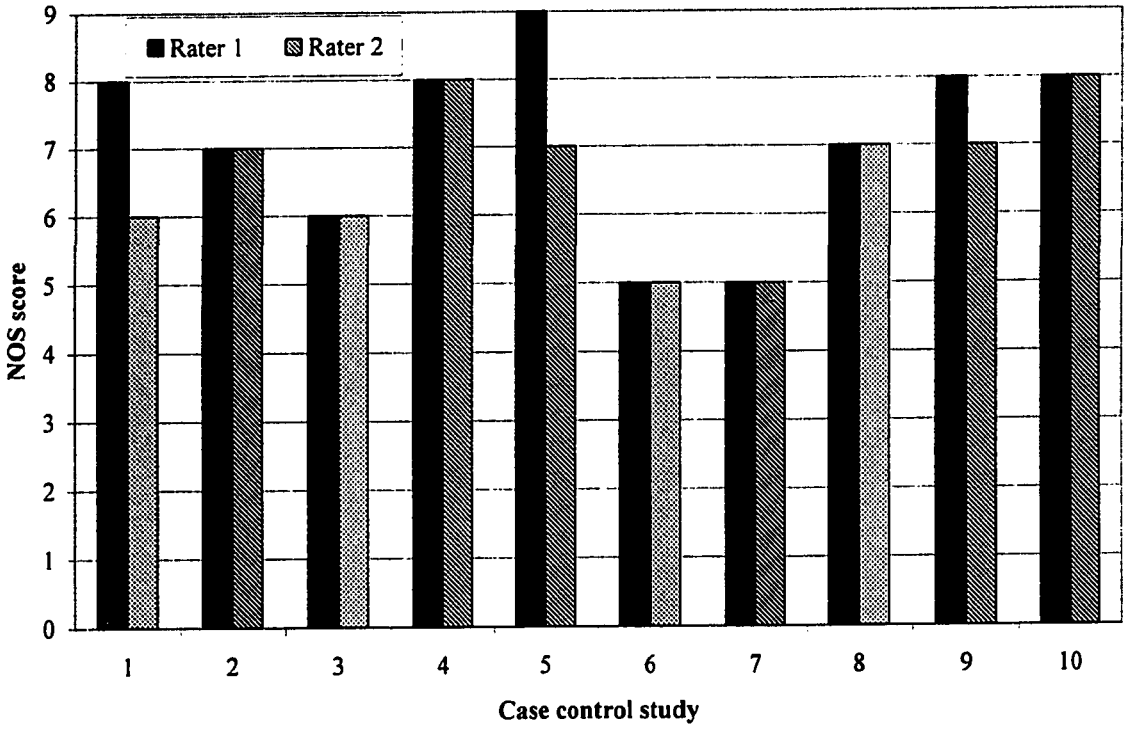


Figure 4. Inter-rater reliability scores for cohort studies

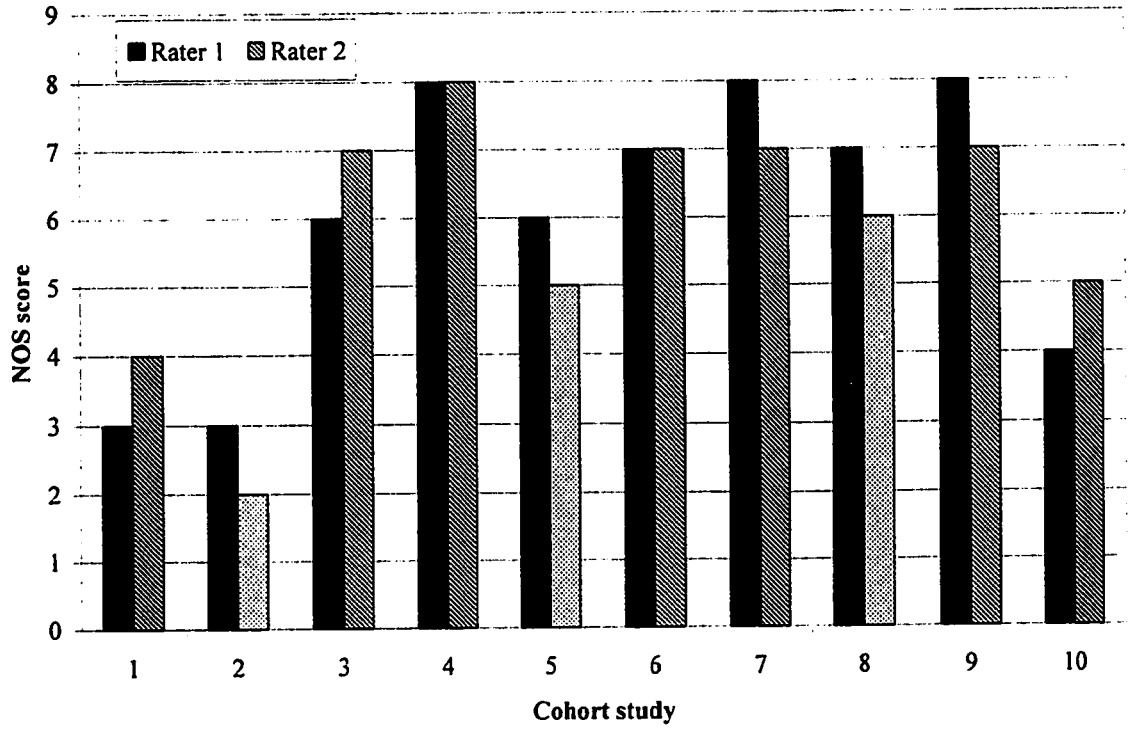


Figure 5: Measure of burden: a comparison of evaluation times for case control studies using NOS and B&D

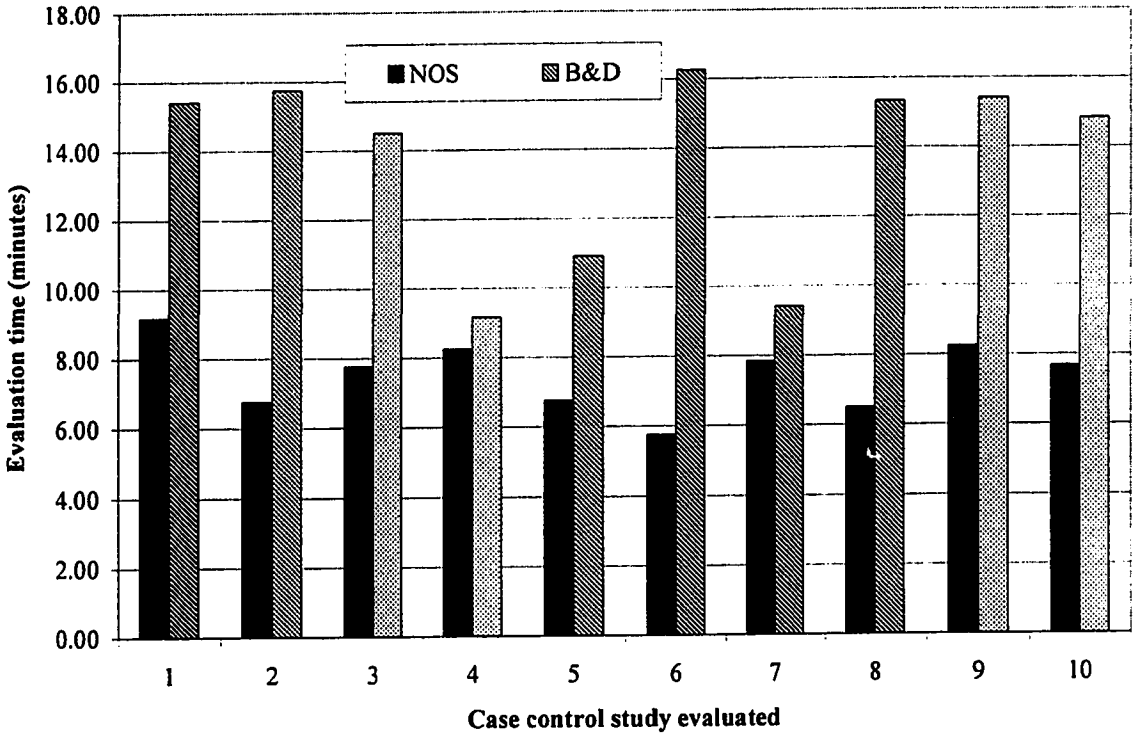
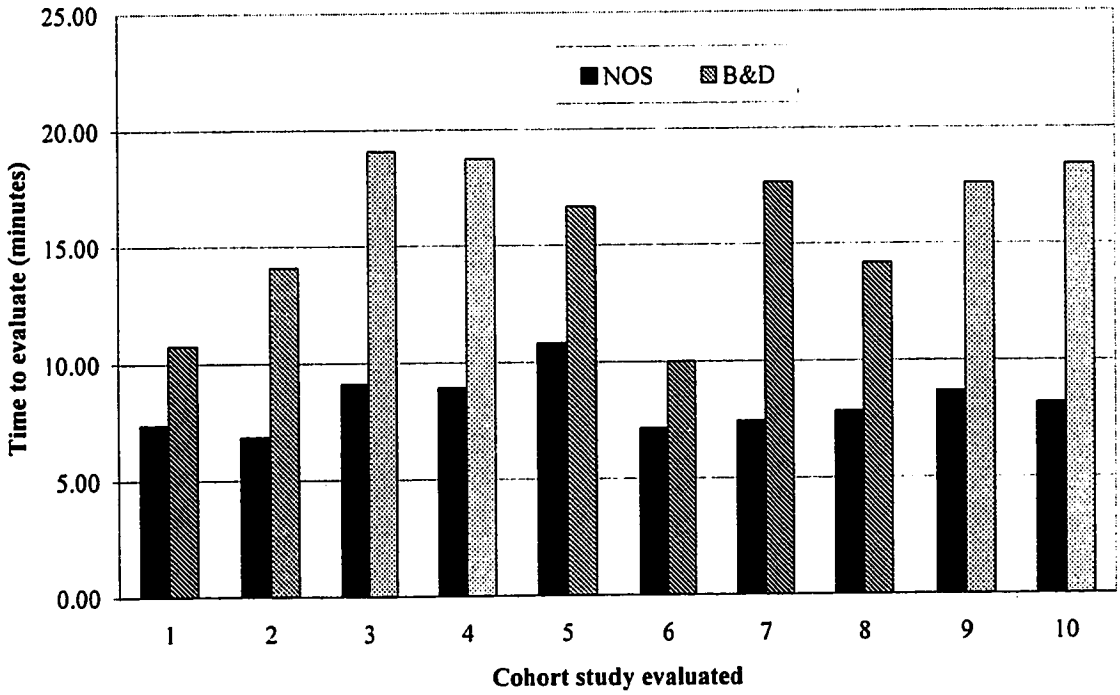


Figure 6: Measure of Burden: a comparison of evalutation times for cohort studies using NOS and B&D



Appendix C Strategies used to search the various electronic data bases

C-1 Strategy used for Medline/PreMedline

1. Hormone replacement therapy/
2. Estrogen replacement therapy/
3. (estrogen$.tw. or oestrogen$.tw.)
4. exp estrogens, conjugated/
5. estrogens, synthetic.mp.
6. (estrogen progestin adj2 therap$).tw.
7. (oestrogen progestin adj2 therap$).tw.
8. or/1-7
9. exp breast neoplasms/
10. (breast adj2(neoplasm$ or cancer$)).tw.
11. breast$.tw.
12. (tumo?r or cancer$ or neoplasm$).tw.
13. or/9-12
14. 8 and 13
15. cohort studies/
16. cohort.tw.
17. relative risk.tw.
18. case-control studies/
19. case-control$.tw.
20. odds ratio$.tw.
21. epidemiologic methods/
22. (causation or causal$).tw.
23. Randomized controlled trials/
24. Or 15-23
25. 14 and 24

C-2 Strategy used to search Current Contents, Dissertation Abstracts and Biological Abstracts

1. (breast adj2 (neoplasm$ or cancer$)).tw.
2. breast$.tw.
3. (tumo?r or cancer$ or neoplasm$).tw.
4. breast cancer
5. or/1-4
6. hormone replacement therapy/
7. estrogen replacement therapy/
8. ((oestrogen or estrogen) and replacement$).
9. (estrogen$ or oestrogen).tw.
10. ((Postmenopaus$ adj hormon$) and replacement
11. (hormon$ and replacement$).tw.
12. or/6-11
13. 5 and 12
14. cohort studies/
15. cohort.tw.
16. relative risk.tw.
17. case-control studies/
18. case-control$.tw.
19. odds ratio$.tw.
20. epidemiologic methods/
21. (causation or causal$).tw.
22. randomized controlled trials/
23. or/14-22
24. 13 and 23

C-3 Strategy used to search Embase

1. estrogen therapy/
2. exp gestagen/
2. 1 and 2
3. (estrogen progestin adj2 therap$).tw.
4. (oestrogen progestin adj2 therap$).tw.
5. or/3-5
6. exp breast cancer/
7. (breast cancer$ or breast neoplasms$).tw.
8. (breast tumor$ or breast tumour$).tw.
9. or/7-9
10. 6 and 10
11. exp risk
12. risk.tw.
13. 12 or 13
14. 11 and 14
15. postmenopause/
16. (post menopaus$ or postmenopaus$).tw.
17. 16 or 17
18. 15 and 18
19. controlled study.de.
20. clinical trial.de.
21. randomized controlled trial.de.
22. clinical trial/
23. cohort analysis/
24. case control study/
25. (cohort or case control$).tw.
26. or/20-26
27. 19 and 27

Appendix D-1 Extraction form for Case Control study **Review:**

AUTHOR(S) ,YEAR:	LANGUAGE
Title:	
Location of study:	Year(s) of study
Aim of Study:	
Outcome:	Criteria:
Exposure:	Formulation of EPT:
Case definition: ☐ Excluded cases with diagnosis prior to exposure	Control definition ☐ Matched according to _____ ☐ Hospital controls ☐ Community controls ☐ Randomly Selected ☐ Excluded History of Disease ☐ Unmatched ☐ Single Control ☐ Multiple Controls_____
Classification of exposure: (current/ever/never)	

Information	CASES	CONTROLS
# identified		
Number excluded and explanation		
Group N		
Group Determination	☐ Direct measurement ☐ Medical records ☐ Self report ☐ Other:	☐ Direct measurement ☐ Medical records ☐ Self report ☐ Other:
Ascertainment Of exposure	☐ Direct measurement ☐ Medical records ☐ Self report Blinding ☐ yes ☐ no ☐ N/A	☐ Direct measurement ☐ Medical records ☐ Self report: Blinding ☐ Yes ☐ No ☐ N/A

Subject Characteristics

CHARACTERISTIC	CASES	CONTROL

Group differences

Comments/concerns:

Data Analysis for : ☐ All subjects ☐ Excludes ascertainment exposure unknown

EXPOSURE	N			CRUDE ODDS RATIO (CI)	ADJUSTED ODDS RATIO (CI)	FACTORS CONTROLLED (in analysis)	TRANSLATION
Estrogen and progestins (ever use) Never use		Cases	Control		Estimation method:	☐	θ = ω = LCL = UCL = SE =
	E						
	$\bar{E}$						
	Total						
		Cases	Control				ϑ = ω = LCL = UCL = SE =
	E						
	$\bar{E}$						
	Total						

Critical issues concerning differences among study groups, study methods, uncertainties regarding data, missing information:

Appendix D-2 Extraction form for Cohort Study: **Reviewer:**

AUTHOR(S) AND YEAR:	LANGUAGE:
Title:	
Location of study:	Dates of enrolment and follow-up:
Endpoint:	Criteria:
Formulation:	Classification of exposure: current/ever/never
Exposure:	Study name (if applicable):
☐ Excluded ♀ with a history of BC	

Methodology : Design

☐ Prospective Cohort ☐ Retrospective Cohort ☐ Internal Cohort ☐ External Cohort

Cohort	Exposed (cohort)	Unexposed
# Identified (describe)		
# Excluded/Lost to Follow-up (describe)		
Group N		
Group selection		
Group Determination	☐ Direct Measurement ☐ Medical Records ☐ Self Report Blinding ☐ Yes ☐ No ☐ NA	☐ Direct Measurement ☐ Medical Records ☐ Self Report Blinding ☐ Yes ☐ No ☐ NA
Ascertainment of Outcome	☐ Direct Measurement ☐ Medical Records ☐ Self Report Blinding ☐ Yes ☐ No ☐ NA	☐ Direct Measurement ☐ Medical Records ☐ Self Report Blinding ☐ Yes ☐ No ☐ NA

Subject Characteristics

CHARACTERISTIC	COHORT	EXPOSED	UNEXPOSED

Critical issues concerning differences among study groups, study methods, uncertainties regarding data, missing information:

Comments/concerns:

Data Analysis: ☐ All subjects ☐ Loss to follow up excluded

OUTCOME EXPOSURE	N			Crude RR (CI)	Adjusted RR (CI)	FACTORS ADJUSTED	TRANSLATION
Outcome Exposure		Cases	Controls			☐ _____ ☐ _____ ☐ _____ ☐ _____ ☐ _____ ☐ _____ ☐ _____ ☐ _____	
	E						
	$\bar{E}$						
	Σ						
Outcome: Exposure:		Cases	Controls			☐ _____ ☐ _____ ☐ _____ ☐ _____ ☐ _____	
	E						
	$\bar{E}$						
	Σ						

Comments/concerns:

Apppendix E Excel calculation spreadsheet used to determine summary estimate for meta-analysis

Study/year	OR	Lower CI	Upper CI	ln of OR	ln of LCI	ln of UCI	UCI - LCI	Std Error	Weight	Sum WY	(Y) x (Y)	Chi Squ T	Chi Squ H
Kaufman 1991	1.7	0.9	3.6	0.5306	-0.1054	1.2809	1.3863	0.3536	7.9958	4.2428	0.282	2.251	0.2566
Palmer 1991	0.6	0.2	2	-0.5108	-1.6094	0.6931	2.3026	0.5874	2.8983	-1.4805	0.261	0.756	2.1552
Yang 1992	1.2	0.6	2.2	0.1823	-0.5108	0.7885	1.2993	0.3314	9.1026	1.6596	0.033	0.303	0.2605
Ewertz 1993	1.36	0.98	1.87	0.3075	-0.0202	0.6259	0.6461	0.1648	36.8059	11.3172	0.0945	3.4799	0.0713
La Vecchia 1995	1.6	0.4	6.3	0.4700	-0.9163	1.8405	2.7568	0.7033	2.0218	0.9503	0.221	0.447	0.0284
Stanford 1995	0.90	0.60	1.20	-0.1054	-0.5108	0.1823	0.6931	0.1768	31.9831	-3.3698	0.0111	0.3550	6.6756
Persson 1997	1.4	0.9	2.2	0.3365	-0.1054	0.7885	0.8938	0.2280	19.2342	6.4718	0.113	2.178	0.0043
Newcomb 1998	1.44	1.19	1.75	0.3646	0.1740	0.5596	0.3857	0.0984	103.3136	37.6726	0.133	13.737	0.0178
Chen 1999	1.51	1.09	2.08	0.4121	0.0862	0.7324	0.6462	0.1648	36.8003	15.1658	0.1698	6.2500	0.1352
Kirsh 2001	1.22	0.72	2.06	0.1989	-0.3285	0.7227	1.0512	0.2682	13.9057	2.7652	0.040	0.550	0.3240
Magnussen 1999	1.63	1.37	1.94	0.4886	0.3148	0.6627	0.3479	0.0887	126.9756	62.0377	0.2387	30.3104	2.3860
Summary								Totals:	391.0369	137.4326	1.5966	60.6166	12.3149
Overall measure	1.4211	1.29	1.57	0.3515	0.2523	0.4506		0.0506					
Chi squ Total	60.6166												
Chi squ Association	48.3017												
Chi squ Homogeneity	12.3149												