DEBATE—continued

Issues to debate on the Women’s Health Initiative (WHI) study. Hormone replacement therapy: an epidemiological dilemma?

K.Machens1,3 and K.Schmidt-Gollwitzer1,2

1Research and Development, Schering AG, Müllerstrasse 170–178, 13342, Berlin and 2Medizinischer Fachbereich, Humboldt-Universität, Berlin, Germany

3To whom correspondence should be addressed. E-mail: kathrin.machens@schering.de

In July 2002 the data of the prematurely stopped Estrogen plus Progestin study of the Women’s Health Initiative (WHI) were presented in the Journal of the American Medical Association. The results of the Heart and Estrogen/Progestin Replacement Study (HERS/HERS II) were published in the same issue. The results of WHI for healthy postmenopausal women often are interpreted to be in analogy with the HERS/HERS II results for postmenopausal women with established coronary heart disease. As a result of HERS/HERS II and WHI, use of HRT in general became questionable. This is an unjustified judgement of HRT in general. This synoptic review and criticism of both studies will show the methodological weaknesses and their consequences and the reasons for limited generalizability of the study results from WHI and HERS/HERS II on normal HRT users.

Key words: disease outcomes/epidemiology/HERS-HERS II/hormone replacement therapy/WHI

Introduction

The benefits of HRT for the prevention of age-related diseases of women are borne out by numerous observational studies, which, together, encompass several hundred thousand woman-years (Grady et al., 1992; Grodstein et al., 1996; Nevitt et al., 1996; Johnson, 1998; Panidis and Stergioupolus, 1998; Nanda et al., 1999; Grodstein et al., 2001; de Kleijn et al., 2002).

However, two of the biggest randomized controlled US studies, the Heart and Estrogen/Progestin Replacement Study or HERS/HERS II (Hulley et al., 1998; Grady et al., 2002; Hulley et al., 2002) and the Women’s Health Initiative Study or WHI (Writing Group for the Women’s Health Initiative Investigators, 2002), failed to demonstrate any protective effect on, for example, the heart in the view of the investigators. Both studies had a randomized, double-blind, placebo-controlled design that promised highly objective results. The medication was the same in both studies: Prempro® (0.625 mg conjugated estrogens and 2.5 mg medroxyprogesterone acetate per day) or placebo. The difference between the studies lay in the population of women chosen for investigation: the HERS and HERS II study looked at women with established heart disease, whereas WHI looked at healthy (at least in the investigators’ view), postmenopausal women.

The objective of the HERS/HERS II study was to ascertain the risk reduction of coronary heart disease (CHD) events; the objective of the WHI study was to investigate the effect on the risk of CHD and other cardiovascular diseases, and as a possible side-effect, an increase in the risk of developing breast cancer. Hip and other fractures were secondary endpoints.

The WHI study includes other investigations, such as the effect of diet and calcium/vitamin D supplementation on the risk of breast cancer and the risk of fracture. These investigations are still ongoing, as is the study of estrogen monotherapy (Premarin®) versus placebo. On May 31, 2002, the investigation of the combined conjugated estrogen/progestin therapy (Prempro®) versus placebo was discontinued prematurely.

The reasons given for the sudden discontinuation of the estrogen plus progestin study were a perceived lack of cardiovascular benefit and a perceived increase in breast cancer, strokes and coronary events (Writing Group for the Women’s Health Initiative Investigators, 2002).

The results of the HERS/HERS II and WHI studies seem to cast doubt on HRT in general and especially long term HRT. The extrapolation of these results on all HRT products apart from their ingredients and administration route, and on all postmenopausal women apart from their age and health status, has stirred concern and anxiety among the medical community and the general population.

In the following we will embark on a comparative analysis of the HERS/HERS II and WHI trials and briefly summarize the conclusions to be drawn from these studies.
**Table I. Characteristics of the Heart and Estrogen/Progestin Replacement Study (HERS), HERS II and Women’s Health Initiative (WHI)**

<table>
<thead>
<tr>
<th></th>
<th>HERS</th>
<th>HERS II</th>
<th>WHI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study design</td>
<td>Randomized, blinded, placebo-controlled</td>
<td>Randomized, open, placebo-controlled</td>
<td>Randomized, blinded, placebo-controlled</td>
</tr>
<tr>
<td>Study objective</td>
<td>Secondary prevention of coronary heart disease</td>
<td></td>
<td>Primary prevention of coronary heart disease; invasive breast cancer as primary adverse outcome</td>
</tr>
<tr>
<td>Study sites</td>
<td>USA (20 centres)</td>
<td>USA (20 centres)</td>
<td>USA (40 centres)</td>
</tr>
<tr>
<td>Follow-up (average)</td>
<td>4.1 years</td>
<td>2.7 years (planned 4 years)</td>
<td>5.2 years (planned 10 years)</td>
</tr>
<tr>
<td>Σ 6.8 years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study population</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No.</td>
<td>2763</td>
<td>2321</td>
<td>16 608</td>
</tr>
<tr>
<td>Women enrolled relative to women screened (%)</td>
<td>4.0</td>
<td>3.7</td>
<td>4.5</td>
</tr>
<tr>
<td>Age*</td>
<td>66.7 years (44–79)</td>
<td>67 years (&lt; 80)</td>
<td>63.3 years (50–79)</td>
</tr>
<tr>
<td>History</td>
<td>Postmenopausal women with established coronary heart diseaseb</td>
<td>None (healthy postmenopausal women)</td>
<td></td>
</tr>
<tr>
<td>Study medication</td>
<td>0.625 mg CEE plus 2.5 mg MPA (Prempro®) 1 tablet/day or placebo</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Age at screening.

bHistory of either: myocardial infarction, coronary artery bypass graft, percutaneous transluminal coronary angioplasty or coronary angiography demonstrating >50% luminal diameter narrowing of a major vessel.

CEE = conjugated equine estrogens; MPA = medroxyprogesterone acetate.

**Study design issues**

**Randomization**

Upon termination, the planned study design of the HERS/HERS II and WHI studies was weakened by excessive drop-out rates, especially in study years 4 and 5, and differential unblinding. This may have resulted in a loss of at least some of the strengths inherent in the randomized, double-blind, placebo-controlled clinical trial design.

An excessive drop-out rate can bring a clinical study to a premature end. In such a setting the prespecified study objectives are often missed as the allowances made for drop-outs during the planning stage of the trial are too small to compensate for excessive loss of patients that may emerge during the trial (Bock, 1998).

The estimated drop-out rate for the HERS study was 2% per year; no estimates are available for the HERS II study. There was a considerable loss of patients in the HERS/HERS II study between years 6 and 7. Almost two-thirds of the patients who had participated in the study for 6 years did not continue through year 7. This holds true for the women in both the HRT and the placebo group. Of the ~1300 women in each group at the inception of the study, some 350 in each group were still active participants in year 7. With this excessive loss of patients continuing, hardly any women would have been left by the end of year 7 if the HERS II study had been conducted over 4 years as originally planned. In the light of these findings the authors’ plea of a lack of favourable cardiovascular effects sounds more like an excuse for being unable to adhere to the study protocol for the planned duration of the trial.

A similar loss of active participants was observed in the WHI study. According to the study design (WHI Study Group, 1998), the drop-out rate was 30% in 9 years. In actual fact, the proportion of women who stopped actively participating in the WHI study was 42% in the HRT and 38% in the placebo group after the mean study duration of 5.2 years. In years 1 to 3, <100 women per year discontinued the study medication. Subsequently, this figure rose to >300 women in year 4, to almost 2000 women in year 5, to >800 women in the HRT and to >1300 women in the placebo group in year 6.

In the WHI trial the conduct of the study was severely compromised by differential unblinding that involved 40.5% of HRT and 6.8% of placebo users (Writing Group for the Women’s Health Initiative Investigators, 2002). Vaginal bleeding, a known side-effect of HRT, was the leading cause of prematurely breaking the blind in the WHI study. Unblinding may alert both patients and investigators to preconceived side-effects of HRT such as deep vein thrombosis, pulmonary embolism, myocardial infarction and stroke, especially in their mild forms, and breast cancer. This heightened index of suspicion (information bias) could have triggered diagnostic measures for these conditions preferentially in known HRT users.

**Study population**

Selecting a suitable study population is one of the most crucial decisions which has major implications on the results of the trial and on the generalizability of the data to the general population. To this end the demographics of the targeted study population are of paramount importance when interpreting study results.

As shown in Table I, only a fraction of the screened women were eventually enrolled: 4.0% (HERS study) (Hulley et al., 1998), 3.7% (93% of 4.0%; HERS II study) (Hulley et al., 2002), and 4.5% (WHI study) (Rossouw et al., 1995). In the HERS/HERS II study, mean patient age was 67 (range 44–79); ~75% of the women enrolled in the study were first users of hormones. In the WHI study, mean patient age was 63 years (range 50–79) at the time of screening; 75% of these women had never taken hormones prior to the study (Table II). Although the age of the study population varies to some extent, the majority of participants were elderly postmenopausal women who had gone through menopause ~10 years before.
In the general population, HRT is usually started in menopausal women who are much younger and presumably healthier than the women enrolled in the HERS/HERS II and WHI studies (Gambacciani et al., 2002).

The WHI study was carried out in postmenopausal women who had a negative history of acute cardiovascular events. In clinically healthy women, asymptomatic vascular disease cannot automatically be ruled out at >60 years of age. In the WHI study, however, 66% of the participants were aged >60 years. The health state of individuals is often reflected by their concomitant medications: in the WHI study there were more ‘healthy’ postmenopausal women on antihypertensive drugs than in the HERS/HERS II study (36 versus 33%); 20% of the ‘healthy’ women in the WHI study were administered aspirin, and 7% were taking statins at study onset. These data indicate that a substantial proportion of the seemingly ‘healthy’ postmenopausal women in the WHI study were on active treatment of cardiovascular risk factors or conditions. This observation does not agree with the authors’ classification of these women as healthy.

Data collection

When the blinded HERS trial was converted into an open, i.e. unblinded, follow-up study (HERS II) there was an abrupt change in the way data were collected from study participants. In the HERS study, the women were seen at 20 centres on a 4 monthly basis. On that occasion patient compliance was checked and information on primary and secondary endpoints was systematically gathered by questionnaire from all patients. In the open HERS II follow-up study the women were regularly contacted by phone every 4 months, using the same questionnaire. In a mean of 32 phone interviews, information was collected from only 62% of HRT and 61% of placebo users, as almost 40% of the study participants did not respond. It was only in the final phone interview that a response rate of 99% was reached. The way in which data are collected by face-to-face questioning or phone interview has a major impact on the type of information gathered and thus on the calculated relative risks associated with the respective conditions. The switch from a favourable trend for cardiovascular events in the blinded HERS trial to an adverse trend in the open HERS II follow-up study may be wholly explained by this abrupt change in data collection.

Biostatistics

A number of criticisms must be raised in relation to the biostatistical analyses. The HERS/HERS II and WHI studies were multicentre trials involving 20 and 40 study centres, respectively. Results can vary considerably between individual centres, a phenomenon known as ‘random error’ or ‘centre effect’. This effect is produced by differences in local patient demographics and patient care. Low patient numbers and event rates can preclude evaluation of the ‘centre effect’ due to loss of statistical power. Nevertheless, stratification of study results by individual centres would have been appreciated.

In the WHI study, interim analyses were performed on a 6 monthly basis. In the course of the trial a total of >70 statistical tests were run. Multiple comparisons of related outcomes do not require correction for multiple testing. Such a correction is needed when measuring unrelated outcomes, e.g. breast cancer and CHD. Failure to adjust for multiple testing will on average produce 1 false positive result in 20 tests at the 5% significance level (Concato et al., 1993). The decision to terminate the estrogen plus progestin part of the WHI study was based on the nominal, i.e. unadjusted, relative risks of breast cancer (HR = 1.26; 95% CI 1.0–1.6) and cardiovascular events (HR = 1.29; 95% CI 1.02–1.63). When the number of statistical tests is taken into account, i.e. adjusted for multiple testing, all the
significantly test results other than those for deep venous thrombosis and pulmonary embolism are no longer significant. In summary, the minor, non-significant increase in relative risks in many conditions can be explained by chance alone.

Clinical endpoints

Cardiovascular events

A comparative analysis of HERS/HERS II and WHI in respect to cardiovascular endpoints can only be made for ‘CHD (non-fatal myocardial infarction and CHD death)’ which was the prespecified common primary endpoint in both studies (Figure 1). For secondary endpoints such as hospitalization for unstable angina pectoris or revascularization procedures (coronary artery bypass graft, CAGB; percutaneous transluminal coronary angioplasty, PTCA), different case definitions or categorizations were used or the respective events were not accounted for in one or the other study.

Since risk estimates for stroke were not broken down by year in the HERS/HERS II trial but only for three time points (HERS = 4.1 years; HERS II = 2.7 years, representing years 5 and 6; HERS/HERS II overall = 6.8 years), these estimates are not truly comparable between the studies.

Coronary heart disease

In the HERS/HERS II trial, risk estimates were similar for the ‘intention to treat’ (analysis by randomization status irrespective of actual treatment) and ‘as treated’ analyses (Table III). After year 1, risk estimates which were 1.52 (non-significant) in both analyses fell progressively to 0.60 (P < 0.05) and 0.58 (non-significant) by year 4, the end of the double-blind randomized HERS study. The conversion from HERS to the open label extension part (HERS II) was associated with an initial increase of the risk estimates to 1.09 and 1.32 (both non-significant) which subsequently fell to 0.99 and 0.71 (both non-significant) at year 6 and after.

For the WHI study, the ‘as treated’ data are unavailable, as are confidence intervals (CI) or P-values (which precludes exclusion of chance as an alternative explanation). It is conspicuous that the WHI risk estimates started at 1.78 at
year 1 and fell progressively to 0.78 at year 6 and later (Figure 1). The only exception to this trend was the risk estimate of 2.38 at year 5 that coincided with a >6-fold surge in the annual drop-out rate relative to the preceding year. The enormous drop-out rate of 42% of participants in the HRT group at the conclusion of WHI is likely to have resulted in a markedly reduced number of observations for the respective endpoints. With decreasing numbers of participants and observations, statistical significance is often lost.

The HERS/HERS II and WHI authors’ conclusion that HRT was ineffective in terms of cardiovascular protection, at least in their trial populations, does not contradict the recognition that HRT has beneficial cardiovascular properties as shown in the Nurses’ Health Study (Grodstein et al., 1996; 2001) and a recent scientific review (Nelson et al., 2002). Current hormone use significantly reduced the risk for CHD and mortality from CHD by 20 and 38% respectively. Experimental data (Clarkson, 2002; Clarkson et al., 2002) are suggestive of a cardioprotective effect by hormones especially in incipient atherosclerosis. In keeping with these findings, HRT was never meant to work in full-blown atherosclerosis.

**Non-cardiovascular events**

The risk estimates of venous thromboembolism (VTE) and total mortality in HERS/HERS II and WHI are comparable in their trends, while those of hip fractures differ markedly, and those of breast cancer only slightly.

**Venous thromboembolism**

During the HERS/HERS II study, the risk estimates of VTE were markedly increased in women on HRT relative to placebo (Figure 3). In year 1, the increase in risk estimates was significant (RH = 3.28; 95% CI 1.1–10.1). Subsequently this increase was progressively reduced through years 2–4: RH = 4.09; 95% CI 0.9–19.3; RH = 2.39; 95% CI 0.6–9.3; and RH = 2.05; 95% CI 0.5–8.2. The wide CI attest to the rarity of VTE events. It is noteworthy that the elevated risk estimates, except for year 1, were non-significant and thus consistent with chance.

In the WHI study of ‘healthy’ postmenopausal women, no CI were provided by study year for any risk estimate. As in HERS/HERS II, the risk estimates of VTE were elevated in the first 5 years (1.67–3.60) in women on HRT but this effect was lost after year 6 (HR = 0.90) (Figure 3).

A substantial fraction of women enrolled in HERS/HERS II and WHI exhibited risk factors for VTE such as advanced age, obesity, smoking or past history of myocardial infarction. Some of these are recognized contraindications for HRT, and most physicians would be hesitant to prescribe HRT to these women. No increased risk of VTE was seen in the younger and healthier population that was studied in the pivotal clinical HRT trials. Because of this beneficial safety profile, HRT is intended for healthier women who are >15 years younger than those in HERS/HERS II (Genazzani, 2002a) and WHI (Genazzani, 2002b). Therefore, it is doubtful whether the HERS/HERS II and WHI results obtained in an elderly population can be generalized to the general population of HRT users.

**Hip fracture**

In HERS/HERS II, risk estimates of hip fracture are provided that are not stratified by personal or familial history of osteoporosis, and only as overall end results of the various trial
phases. At the conclusion of the blinded randomized HERS study after 4.1 years, the risk estimate was 1.16 (non-significant; 95% CI 0.5–2.0), which contrasted with the risk estimate of 2.11 (significant; 95% CI 1.1–4.2) that was obtained 2.7 years thereafter when the open label extension HERS II study was terminated (Hulley et al., 2002). The overall combined risk for HERS/HERS II was calculated as 1.61 (marginally significant; 95% CI 1.0–2.7) (Figure 4).

In the WHI study, risk estimates of a hip fracture were reduced from study inception to 0.64 and varied between 0.87 and 0.55 in the subsequent years (Figure 4). Again no CI were given and thus statistical significance of these estimates cannot be assessed. The overall WHI results revealed a significant risk reduction using the nominal (unadjusted) 95% CI (HR = 0.66; 95% CI 0.45–0.98). As with other significant results, this significance is lost after adjustment for multiple testing (HR = 0.66; adjusted 95% CI 0.33–1.33).

This non-significant risk reduction of hip fractures does not agree with the results of a recent meta-analysis (Nelson et al., 2002) that found a risk reduction of fractures at different sites (non-vertebral fractures, hip fractures, wrist fractures, vertebral fractures) with HRT use.

**Total mortality**

Risk estimates of total mortality were only yielded for the entire duration of the HERS/HERS II study which was 6.8 years. There was no significant difference between the HRT and the placebo group (RH = 1.10; 95% CI 0.92–1.31).

In WHI, the HRT and placebo group likewise did not differ in total mortality except for year 1 (HR = 1.24). The absence of CI for total mortality over time does not allow for an evaluation of statistical significance.

**Malignancies**

The disease outcomes of breast, colorectal and endometrial carcinoma in HERS/HERS II and WHI were comparable. In HERS/HERS II, risk estimates were only calculated for the entire duration of HERS (4.1 years), HERS II (2.7 years), and HERS/HERS II combined (6.8 years).

**Breast cancer**

At the end of the HERS study after a mean of 4.1 years, risk estimates of breast cancer were slightly, although non-significantly, increased (RH = 1.38; 95% CI 0.82–2.31). When the open HERS II study ended after a mean of another 2.7 years,
these risk estimates were further reduced ($RH = 1.08$; 95% CI $0.52$–$2.24$) and still remained non-significant. The overall risk estimates for the entire duration of HERS/HERS II of 6.8 years were calculated at $1.27$ (non-significant; 95% CI $0.84$–$1.94$) (Figure 5). Similar but even smaller risk estimates were obtained in the ‘as treated’ analysis at the conclusion of the HERS/HERS II study ($RH = 1.11$; 95% CI $0.61$–$2.03$) (Table IV).

For the WHI study, no CI and no ‘as treated’ analysis was provided. Therefore the significance of the risk estimates cannot be determined. Risk estimates of invasive breast cancer were reduced from the onset of the WHI trial ($HR = 0.62$ and HR $= 0.83$ in year 1 and 2, respectively) and continued to rise through year 5 ($1.16$–$2.64$) before falling to $1.12$ after year 6 (Figure 5). There was no increased breast cancer risk in women who had never used postmenopausal hormones before WHI compared with women on placebo ($HR = 1.06$; 95% CI $0.81$–$1.38$). With regard to in situ breast cancer, there was no difference between the HRT and placebo women.

Assuming a temporal sequence from in situ to invasive breast cancer, the rate of in situ breast cancer should have risen first in the early years of HRT use before the rate of invasive breast cancer followed suit. This increase should have been even more pronounced when one assumes that HRT promotes the growth of microscopic tumour foci that are already initiated before the start of HRT. No difference in risk estimates was seen for in situ or invasive breast cancer in the first 2 years. All observations made in HERS/HERS II and WHI are totally consistent with chance. It is well recognized that invasive breast cancer takes $>10$ years to evolve from microscopic foci to tumours that are detectable (von Fournier et al., 1980). The women in the HERS/HERS II and WHI studies were administered HRT for a mean of 6.8 and 5.2 years respectively. Such a time span obviously is too short for the evolution of invasive breast cancers.

Colorectal carcinoma

The risk for colorectal carcinoma was non-significantly reduced in HERS after a mean of 4.1 years ($RH = 0.69$; 95% CI $0.3$–$1.5$), in HERS II after a mean of 2.7 years ($RH = 1.01$; 95% CI $0.4$–$2.4$), and in HERS/HERS II after a mean of 6.8 years ($RH = 0.81$; 95% CI $0.46$–$1.45$). In the actual treatment
‘as treated’) analysis HERS/HERS II risk estimates were somewhat lower, albeit still non-significant (RH = 0.58; 95% CI 0.25–1.35).

For the WHI study, the ‘as treated’ data are unavailable, as are CI. During the entire duration of the WHI study, risk estimates of colorectal carcinoma, except for year 2 (HR = 1.17), were appreciably reduced in HRT users (0.72–0.47).

Endometrial carcinoma

A similar pattern was seen in endometrial carcinoma. Risk estimates were non-significantly reduced in HRT users after a mean of 4.1 years in HERS (RH = 0.39; 95% CI 0.1–2.0). In the HERS II extension study, not a single case of endometrial carcinoma was encountered in women on HRT; therefore no risk estimates could be calculated for HERS II. After a mean of 6.8 years, the combined HERS/HERS II risk estimates were reduced even further (RH = 0.25; 95% CI 0.1–1.2) without reaching statistical significance.

In the WHI study, risk estimates of endometrial carcinoma fell progressively from year 1 to year 6 (HR = 0.95–0.17) with the exception of year 4 (HR = 1.91). Although these data may hint at an endometrial protection afforded by combined HRT (estrogen plus progestin) the statistical significance of this trend is not known due to the unavailability of CI.

Conclusion

Applying epidemiological criteria, the results of the HERS/HERS II and WHI trials are neither as disappointing nor as alarming as portrayed in the mass media. The non-significant results of the HERS/HERS II and WHI studies that were obtained in elderly women obviously cannot be generalized to the general population of postmenopausal women who are 15 years younger and thus more healthy. Nevertheless both trials are a testimony to the tremendous difficulties and obstacles that blinded randomized trials of such a huge scale inevitably face. Planning is one step, the management of the study the other. If both contain weaknesses the stated interpretations of the trial data should be perceived cautiously.

References


Submitted on April 4, 2003; accepted on June 20, 2003