BMJ open

Comparing hormone therapy effects in two RCTs and two large observational studies that used similar methods for comprehensive data collection and outcome assessment

Arthur Hartz,[1] Tao He,[2] Robert Wallace,[3] John Powers[4]

To cite: Hartz A, He T, Wallace R, *et al*. Comparing hormone therapy effects in two RCTs and two large observational studies that used similar methods for comprehensive data collection and outcome assessment. *BMJ Open* 2013;**3**:e002556. doi:10.1136/bmjopen-2013-002556

▸ Prepublication history and additional material for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2013-002556).

Received 8 January 2013
Revised 3 June 2013
Accepted 12 June 2013

For numbered affiliations see end of article.

Correspondence to
Dr Arthur Hartz;
hartzarthur@gmail.com

ABSTRACT

Objectives: Prospective observational studies (OSs) that collect adequate information about confounders can validly assess treatment consequences. However, what constitutes adequate information is unknown. This study investigated whether the extensive information collected by the Women's Health Initiative (WHI) in two OSs and two randomised controlled trials (RCTs) was adequate.

Design: Secondary analysis of WHI data. Cox regression was used to select from all baseline risk factors those that best predicted outcome. Cox regression that included these risk factors was used for two types of analyses: (1) comparing RCT and OS assessments of the effects of hormone therapy on outcome for participants with specific characteristics and (2) evaluating whether adjustment for measured confounders could eliminate outcome differences among datasets.

Setting: The WHI included more than 800 baseline risk factors and outcomes during a median follow-up of 8 years.

Participants: 151 870 postmenopausal women ages 50–79.

Primary and secondary outcome measures: Myocardial infarction and stroke.

Results: RCT and OS results differed for the association of hormone therapy with outcome after adjusting for confounding factors and stratifying on factors that were hypothesised to modulate the effects of hormone therapy (eg, age and time since menopause) or that empirically modulated the effects of hormone therapy in this dataset (eg, blood pressure, previous coronary revascularisation and private medical insurance). Some of the four WHI datasets had significantly worse outcomes than others even after adjusting for risk and stratifying by type of hormone therapy, for example, the risk-adjusted HR for myocardial infarction was 1.37 ($p<0.0001$) in an RCT placebo group compared with an OS group not taking hormone therapy.

Conclusions: Apparently the WHI did not collect sufficient information to give reliable assessments of treatment effects. If the WHI did not collect sufficient data, it is likely that few OSs collect sufficient information.

ARTICLE SUMMARY

Article focus

- Observational studies (OSs) are frequently used to compare outcomes of patients who choose different treatments.
- Results of OSs may be invalid because of confounding due to an association between patient risk and treatment choice.
- The present study assessed whether the extensive information collected by the Women's Health Initiative (WHI) was adequate to eliminate confounding and give valid results.

Key messages

- The effects of hormone therapy on stroke and myocardial infarction differ for OSs and randomised controlled trials even after taking advantage of extensive participant information to remove confounding and to select similar participants.
- Participants who self-selected for different studies had different outcomes that could not be explained by differences in measured risk factors.
- As comprehensive data such as collected by the WHI appear to be inadequate to ensure the validity of an observational study, it is unclear what observational study results can be accepted with confidence.

Strengths and limitations of this study

- The WHI dataset is unusually comprehensive and provided a good test of whether excellent datasets can ensure valid results for an observational study. The conditions for valid OSs were not identified.

Medical practice often depends on observational studies (OSs) that compare outcomes of similar patients treated differently. However, OS results may be erroneous because patient risk factors are confounded with treatment choice. Only if confounding

factors can be adequately measured, can their effects be removed with statistical methods. The success of removing confounding errors has been vigorously debated.[1–3]

The strongest evidence against the validity of the OSs has been discrepancies between OSs and randomised controlled trials (RCTs). In particular, RCTs from the Women's Health Initiative (WHI) found that hormone therapy (HT) increased the risk of myocardial infarction (MI)[4] or had no effect[5] and increased the risk of stroke.[4 5] These findings contradicted a large body of well-performed OSs suggesting that HT may reduce the risk of cardiovascular disease by 30–50%.[6–8]

However, RCT/OS discrepancies do not prove that the OS design is invalid. Another possibility is that the discrepancies are caused by differences in characteristics of the study population, therapy or outcome measurements (eg, duration of follow-up). For example, the women evaluated in the WHI RCT were older than those in most OSs, and there is some evidence that HT has a greater adverse effect on older women or women who began HT several years after menopause.[9–12] There is also evidence that the influence of HT on MI risk is greatest soon after initiation,[13] and OSs that can follow participants soon after they begin therapy may give results similar to RCTs.[10 14] It may be possible that other patient characteristics (eg, obesity, smoking or health status) that differ between types of studies alter the associations between HT and outcomes.

The WHI offers an excellent opportunity to assess the value of OSs for three reasons: (1) The same type of data were collected in almost the same way for two RCTs and two OSs of HT; (2) the data collected included comprehensive information about numerous potentially relevant risk factors that are rarely available in OSs, including many often suspected to cause confounding (eg, those related to socioeconomic status, functional status, psychological status, lifestyle factors and healthcare behaviours) and (3) the sample sizes were large enough to enable subgroup comparisons.

METHODS

The ability of an OS to eliminate confounding was examined by testing three hypotheses:

1. Result differences between OSs and RCTs can be eliminated by adjusting for the WHI risk factors.
2. Differences between OSs and RCTs are caused by differences in modulating factors such as the time after menopause that HT is initiated,[9–12] the time OS participants are on HT prior to beginning the study[13 14] or other participant characteristics that have not been previously suggested.
3. Confounding factors associated with which specific WHI study recruited the participant can be eliminated by adjusting for the WHI risk factors.

WHI dataset

Data were obtained from the WHI, which has been described in detail.[5 6] The study was approved by institutional review boards, and all participants signed informed consent forms. In brief, it was a long-term national health study that focused on strategies for preventing heart disease, breast and colorectal cancer and osteoporosis in postmenopausal women. Women aged 50–79 were enrolled from 1993 to 1998 at 40 clinical centres throughout the USA for clinical trials. Women were asked to enrol in an RCT and those who were not ineligible or not interested were given the opportunity to enrol in the WHI OS.

There were four WHI studies relevant to the present analysis: (1) an RCT of oestrogen therapy (E-alone) for women without a uterus, (2) an RCT of oestrogen plus progesterone (E+P) for women with a uterus, (3) an RCT of diet and (4) WHI OS with no interventions. The RCT of diet served as a second OS for the effects of HT because HT use was not randomised for these patients. Participants who were enrolled in the RCT for diet as well as an RCT for HT were considered to be only in the RCT for HT dataset.

For follow-up and outcome ascertainment all participants completed a self-administered, self-report. This report was completed semiannually by the RCT participants and annually by the OS participants. Adjudicated outcomes were based on medical records, autopsy reports and death certificates.

The more than 800 baseline risk factors analysed in the present study were in the following categories: demographics, general health, clinical and anthropometric, functional status, healthcare behaviours, reproductive, medical history, family history, personal habits, thoughts and feelings, therapeutic class of medication, hormones, supplements and dietary intake.

Statistical analysis

The Cox proportional hazard regression analysis was used to test the association between outcome and the primary risk factor after adjusting for covariables. The outcomes analysed in this study were MI or stroke that developed after the participants were enrolled in the study. The primary risk factors were HT (either the binary variable for any HT use or the three category variable for use of E-alone, E+P or neither) or the categorical variable for the four datasets.

The primary risk factors were represented by an indicator variable for every category except the reference category. The HR associated with an indicator variable for a category represented the risk for participants with that variable compared with the risk of participants in the reference category. The reference category for the HT variables was no HT use, and the reference category for dataset was the WHI OS.

To identify which covariables should be included in a Cox model, we first tested the statistical significance of more than 800 risk factors by including only the risk factor and age in the Cox model for a given outcome. All risk factors that were statistically significant at the $p<0.01$ level after adjusting for age alone were then

included in a backwards stepwise Cox proportional hazard regression analysis, and variables that remained statistically significant at the p<0.0001 level were retained in the model. We then used the Cox forward stepwise procedure to test whether any of the variables not already in the model could enter at the p<0.0001 level. It is unlikely that many of these variables were significant by chance alone and even less likely that adjusting for spurious variables would distort the association between HT and outcome.

To identify which risk factors modulated the association between HT and outcome we tested the interactions of HT with the risk factors that had been tested with the timing hypothesis or that had a statistically significant association with outcome at the p<0.01 level after adjusting for age and dataset.

In an analysis that only included OS participants not taking HT at baseline, follow-up began at the time the participant completed the questionnaire that first reported HT or, if they never began HT, follow-up began at the time they completed their first questionnaire after baseline. (If follow-up for these participants had begun as late as it did for the HT participants, it would have diminished the HR associated with HT.) The baseline age of participants in this analysis was computed for the time that follow-up began.

Stepwise procedures were used to find a logistic regression equation that included the risk factors independently associated at the p<0.0001 level with taking baseline HT in the WHI OS. An individual's propensity score was the probability derived from her characteristics and the estimated parameters in this equation. We evaluated whether grouping participants with similar propensity scores decreased confounding in the OSs so that OS and RCT results became more similar.

The median follow-up time was 8 years. However, for the E+P RCT, treatment was ended after a mean follow-up of 5.2 years even though follow-up on all participants was continued. To make time on HT in the study comparable for the OS and each RCT, we ended follow-up at 5 years.

All statistical analyses were performed using SAS V.9 (SAS Institute Inc, Cary, North Carolina, USA).

Sample size

Participants available for analysis included 161 748 WHI participants: 93 651from the observational study, 16 590 from the RCT of oestrogen plus progesterone (E+P), 10 722 from the RCT of oestrogen only (oestrogen-alone) and 40 785 additional women who were in the diet study and not in an RCT of HT. Of the 161 748 WHI participants, 9584 were excluded because they did not meet the following RCT exclusion criteria: platelets less than 75 000/mm^3, haematocrit less than 32%, oral daily use of a glucocorticosteroid, body mass index less than 18, systolic blood pressure greater than 200 mm Hg, diastolic blood pressure greater than 105 mm Hg, breast cancer ever, other cancers in the last 10 years, or stroke, transient ischaemic attack (TIA) or MI in the past 6 months. An additional 294 were missing information on the use of HT at baseline.

Missing data for the covariables were imputed by the mean value for ordinal or binary variables and the mode value for variables with three or more categories. After determining which risk factors were independently associated with a given outcome at the p<0.0001 level, we created a corresponding indicator variable for each of those risk factors that indicated if the variable was missing. If the missing indicator variable was statistically significant at the p<0.05 level, participants missing the corresponding risk factor were excluded. There were 146 936 participants included in the fully adjusted Cox model for MI and 149 470 included in the fully adjusted Cox model for stroke. The ability of the Cox model to predict outcome as measured by the C statistic was not improved by excluding participants with estimated values of the covariates.

RESULTS

Baseline participant characteristics for participants in the four datasets are compared in table 1. For two datasets participants on HT were compared with participants without HT. That was not necessary, however, for the RCTs for E+P and for E-alone because randomisation in these studies made the treatment arm unrelated to baseline characteristics. In the OS and RCT for diet datasets the risks due to age, race, income, educational level, physical functioning and smoking were most favourable for participants on E+P and least favourable for participants not taking HT. With the exception of smoking these characteristics were also more favourable for participants in the RCT for E+P than in the RCT for E-alone. Both socioeconomic status variables (education and income levels) are lower for the two RCTs of HT datasets than for the other two datasets, p<0.0001. For this reason it was important to evaluate whether socioeconomic status influenced the association between HT and outcome.

Propensity score

The logistic regression equation to predict the probability that a participant in the OS used HT (ie, the propensity score) included 94 independent risk factors statistically significant at the p<0.0001 level and had a C statistic of 0.90, indicating that the equation was highly predictive of HT use.

Risk factors for MI and stroke

We identified 16 risk factors (in addition to dataset) that were independently associated with MI at the p≤0.0001 level. The variables and their associated χ^2 value for the full dataset in parenthesis were age (594.3), taking medication for diabetes (284.3), smoking at baseline (182.4), systolic blood pressure (150.1), history of coronary artery bypass surgery (110.1), history of cardiovascular disease

Table 1 Percentage of participants in a given category by dataset and type of hormone therapy

	WHI OS			RCT for diet			RCTs for HT	
Variables	**E+P**	**E-alone**	**No HT**	**E+P**	**E-alone**	**No HT**	**E+P**	**E-alone**
Sample size	17 618	21 659	44 597	8907	11 880	19 968	16 581	10 719
Age (years)								
≤55	25.6	19.0	13.4	27.1	20.6	14.7	16.6	16.4
>70	9.3	16.8	25.0	5.6	11.5	17.8	17.8	20.1
Race								
Non-white	11.3	15.5	19.8	10.8	15.5	22.4	16.0	24.7
Family income								
<$35 000	23.0	33.4	42.7	23.4	33.9	41.9	45.3	54.5
>$75 000	29.1	20.0	14.5	27.4	18.8	13.7	12.5	8.1
Education level								
≤HS grad	13.2	20.8	24.8	13.7	21.6	23.5	26.1	32.4
Col grad	54.1	38.4	38.1	50.7	35.3	37.0	34.6	23.7
P Funct >75	80.2	68.8	68.2	78.4	67.7	67.4	73.6	61.5
Med visit	84.4	85.3	76.7	85.7	86.0	76.2	68.5	72.3
Smoking								
Past	46.0	42.7	40.1	44.6	42.0	40.2	39.2	38.0
Current	5.1	5.5	7.0	5.5	5.6	6.8	10.3	10.3
Meno sympt	77.3	70.4	64.8	70.9	64.3	59.5	61.9	60.5

All characteristics differed among the four datasets and among treatment groups within the observational study and RCT for diet datasets at the p<0.0001 level.
Col, college; E-alone, oestrogen alone; E+P, oestrogen plus progesterone; grad, graduate; HS, high school; HT, hormone therapy; Med visit, visit to a physician within the past year; Meno sympt, history of menopausal symptoms; OS, observational study; P Funct, physical function score from the SF-36; RCT, randomised control trial.

(67.1), limited in climbing stairs (62.8), worse general health (52.1), family history of MI (50.0), lower income (46.4), current history of MI (44.2), white race (44.1), the ratio of waist circumference to hip circumference (38.1), hypertensive medications (33.4), taking calcium channel blockers (24.0) and higher haematocrit (18.6). The C statistic of the predictive value for this equation was high, 0.78 (95% CI 0.77 to 0.79).

Twelve risk factors were independently associated with stroke at the p≤0.0001 level: age (667.4), systolic blood pressure (181.4), history of diabetes (110.3), medication for hypertension (85.3), current smoking (79.9), physical function (68.2), history of stroke (49.1), history of cardiovascular disease (38.8), TIAs (30.8), cardiotonic medication, especially digitalis (27.1), lower income (21.7) and lifetime HT duration (14.9). The C statistics for these variables was 0.76 (95% CI 0.76 to 0.77).

Association of HT with MI and stroke

The risk-adjusted HRs for a specific type of HT (E+P or E-alone) and for either HT are shown in table 2 for each dataset. In the WHI OS dataset E+P and E-alone had similar HRs. In the diet dataset E-alone was significantly protective for MI (HR=0.65) but E+P was not (HR=0.96, p=0.04 for the difference between HRs for E-alone and E+P), and there was no association of either type of HT with stroke. In the RCT datasets there was an association of E+P with an increased risk of MI (HR=1.30) as well as stroke (HR=1.34), but E-alone was not associated with MI.

To test for differences in HRs among the datasets, we combined all datasets and included main effects, interactions between HT and dataset and risk factors in the Cox model. The MI HRs for E+P was larger in the E+P RCT than in the OS (p=0.07), and the MI HR for E-alone was higher in the RCT for E-alone than in the diet dataset (p=0.06). For stroke, where the evidence for the HT risk is stronger, the HR in the combined RCT datasets was significantly higher than it was in the WHI OS dataset (p<0.0001) and in the diet dataset (p=0.005).

Influence of patient characteristics on the association between HT and outcomes

The analyses reported in tables 3 and 4 examined how OS and RCT differences might be influenced by the timing of the HT with respect to age, menopausal status and previous hormones. Also these tables show the effects of additional adjustment for confounding using propensity scores. The HRs and their CIs are presented for women on any HT. Where it might be informative, HRs without CIs are presented for women on a specific type of HT (either E+P or E-alone).

Myocardial infarction

Table 3 presents the MI HR for HT, E+P and E-alone. The timing hypothesis suggests that HRs should be significantly lower in the 50–59 age group or in the group with menopause less than 10 years than in the other groups, but none of these differences were significantly different in the expected direction. To the contrary, the

Table 2 Risk-adjusted HRs for hormone therapy in different datasets

		Myocardial infarction		Stroke	
Dataset	**HT type**	**HR**	**95% CI**	**HR**	**95% CI**
WHI OS	Any E	0.83	(0.72 to 0.95)	0.85	(0.70 to 1.03)
	E+P	0.86*	(0.70 to 1.05)	0.82*	(0.65 to 1.04)
	E-alone	0.80†	(0.69 to 0.94)	0.88‡	(0.71 to 1.11)
Diet RCT	Any E	0.75	(0.62 to 0.89)	1.04	(0.80 to 1.37)
	E+P	0.96	(0.75 to 1.22)	1.00	(0.72 to 1.39)
	E-alone	0.65†§	(0.53 to 0.81)	1.07	(0.79 to 1.45)
HT RCT	Any E	1.18	(0.99 to 1.41)	1.29	(1.05 to 1.58)
	E+P	1.30	(1.02 to 1.65)	1.34	(1.02 to 1.77)
	E-alone	1.05	(0.81 to 1.36)	1.23	(0.91 to 1.67)

*Differs from the comparable RCT HR at the p<0.01 level.
†Differs from the comparable RCT HR at the p=0.02 level.
‡Differs from the comparable RCT HR at the p=0.06 level.
§Differs from 1.00 at the p<0.0001 level.
Any E, E+P or E-alone; E-alone, oestrogen alone; E+P, oestrogen plus progesterone; HT, hormone therapy; OS, observational study; RCT, randomised controlled trial; WHI, Women's Health Initiative.

E+P HR for women aged 50–59 was much higher (1.63) than it was for older women (1.01 for women age 60–69).

The HR for HT during the first 3 years (1.26) is greater than the subsequent risk (1.08). For the RCT for E+P the difference is greater, 1.45 vs 1.11, and the test of the time-dependent covariables of duration of exposure was of marginal statistical significance (p<0.05). Since OS participants on HT began HT several years before enrolment, a diminished effect of HT with time could explain an OS/RCT difference. However, results of other analyses do not support this explanation: there was no evidence that previous HT exposure reduced the HR in the RCT (ie, the HR was lower for participants with no previous exposure, 1.07, than for those with previous exposure, 1.51), and there was no indication in the WHI OS dataset of increased MI risk for participants who began HT after study baseline, the HR was lower than it was for participants who began HT at baseline. (Information on HT usage after baseline was not available for the diet RCT study.)

The last rows in table 3 are HRs stratified by propensity scores. Stratifying by propensity score in addition to adjusting for the significant covariables was expected to reduce confounding, but there was no evidence that doing this gave results similar to the RCTs.

Additional factors that significantly modulated the association between HT and MI in the OS dataset at the p<0.05 level included blood pressure, previous coronary revascularisation, hours of sleep, haematocrit, working status, thyroid disease, antineoplastics, private medical insurance, bone fracture after age 55, colon polyps, ever lived or worked on farm and hostility. Neither education nor income was a statistically significant modulating variable. No factors that significantly modulated the HT HR in the WHI OS dataset also significantly modulated this HR in the RCT datasets. The MI HRs in the RCT and OS datasets did not become similar if they were stratified by the modulating variables.

Stroke

Although E+P and E-alone had similar associations for stroke, results in table 4 include only the HRs for HT and no HRs for E+P and E-alone. As shown in this table there was no consistent evidence that the HT HR for stroke was lower for women who were younger or had menopause recently. In contrast to the MI analyses, there was also no RCT evidence that the HT HR for stroke was stronger soon after beginning HT.

The only variable found to significantly influence HT HR for stroke in the WHI OS dataset was endometrial aspiration; the HR was 0.85 for those who had had an endometrial aspiration and 1.16 for participants who did not (p<0.001). Stratifying on this variable did not make the OS and RCT results more similar. In addition, the lack of an obvious medical explanation, the number of factors tested and the lack of this relationship in the RCT datasets makes it more likely that this result occurred by chance.

After recalculating the HR in the WHI OS dataset for only those participants with midrange of propensity scores (those with a probability of using HT between 0.25 and 0.75), the HR for stroke was virtually unchanged. This suggests that adjusting for the propensity score did not diminish confounding.

Adequacy of WHI information to eliminate confounding

In table 5, the MI risks are compared for participants in the four different WHI datasets who are on the same treatment at baseline (E+P, E-alone or no HT). The HR in the table represents the risk of the outcome for participants in that dataset compared with participants on the same treatment in the WHI OS dataset. If the WHI variables are adequate to eliminate confounding, the adjusted HRs should be near 1.00.

Some HRs shown in the table were statistically significant at p<0.0001. For participants not taking HT the risk-adjusted HR was 1.37 for the RCT for E-alone. For

Table 3 MI HRs for hormone therapy in subgroups defined by participant characteristics associated with hormone exposure

Subgroup within dataset	Dataset					
	RCTs for HT		RCT for diet		WHI OS	
	MI HR for HT in the subgroup of the indicated dataset (Numbers in parentheses are HRs for E+P and E-alone)					
		95% CI		95% CI		95% CI
All participants	1.18 (1.30,1.05)	0.99 to 1.41	0.75 (0.95,0.65)	0.62 to 0.89	0.83 (0.86, 0.80)	0.72 to 0.95
Age						
50–59	1.25 (1.63,0.69)	0.80 to 1.96	0.57 (0.73,0.44)	0.37 to 0.89	0.73 (0.74,0.60)	0.54 to 0.99
60–69	1.01 (1.05,0.95)	0.78 to 1.32	0.73 (0.88,0.65)	0.56 to 0.94	0.87 (0.97,0.81)	0.71 to 1.07
70–79	1.46 (1.46,1.20)	0.99 to 2.15	0.87 (1.33,0.74)	0.65 to 1.18	0.84 (0.75,0.86)	0.68 to 1.03
Years since meno						
<10	1.03 (1.14,0.77)	0.73 to 1.46	0.83 (1.01,0.68)	0.60 to 1.15	0.85 (0.94,0.80)	0.73 to 0.99
10–19	0.95 (1.06,0.74)	0.68 to 1.34	0.67 (0.95,0.47)	0.48 to 0.95	0.77 (0.74,0.72)	0.44 to1.35
>19	1.41 (1.77,1.23)	1.08 to 1.85	0.69 (0.60,0.70)	0.50 to 0.93	1.35 (0.61,1.28)	0.46 to 3.96
HT after baseline			No data		0.71 (0.76, 0.72)	0.57 to 0.88
Follow-up for RCT						
End 3 years after enrolment	1.26 (1.45,1.06)	1.00 to 1.58			0.86 (0.87,0.85)	0.71 to 1.02
Begin 3 years after enrolment	1.08 (1.11,1.04)	0.82 to 1.41			0.79 (0.84,0.73)	0.64 to 0.97
Previous use of HT						
No	1.07 (0.96,1.20)	0.86 to 1.32				
Yes	1.51 (1.12,1.46)	1.09 to 2.08				
Propensity score						
<0.25	1.26 (1.27,1.18)	0.96 to 1.66	0.75 (0.76, 0.71)	0.48 to 1.17	0.98 (1.04,0.83)	0.72 to 1.34
0.25–0.75	1.11 (1.32,1.02)	0.87 to 1.42	0.80 (1.01, 0.69)	0.64 to 1.00	0.85 (0.81, 0.86)	0.72 to 1.01
>0.75	1.05 (NA, 0.93)	0.42 to 2.64	0.69 (1.08, 0.67)	0.34 to 1.40	0.76 (0.96, 0.74)	0.45 to 1.27

E-alone, oestrogen alone; E+P, oestrogen plus progestin; HT, hormone therapy; meno, menopause; MI, myocardial infarction; NA, not available because only one MI case in this group; OS, observational study; RCT, randomised controlled trial; WHI, Women's Health Initiative.

Table 4 Stroke HRs for hormone therapy in subgroups defined by participant characteristics associated with hormone exposure

Subgroup within dataset	Dataset					
	RCTs for HT		RCT for diet		WHI OS	
	Stroke HR for HT in the subgroup of the indicated dataset					
		95% CI		95% CI		95% CI
All patients, full follow-up	1.29	(1.05, 1.58)	1.04	(0.80, 1.37)	0.85	(0.70, 1.03)
Age						
50–59	1.03	0.59 to 1.82	0.48	(0.24, 0.98)	1.04	0.61 to 1.78
60–69	1.65	1.20 to 2.27	1.33	0.90, 1.96	0.96	0.71 to 1.29
70–79	1.11	0.81 to 1.51	1.14	0.74, 1.77	0.75	0.55 to 1.01
Years since menopause						
<10	1.33	0.87 to 2.05	0.85	0.52, 1.40	0.79	0.62 to 1.00
10–19	1.4	0.96 to 2.06	1.05	0.61, 1.82	0.67	0.39 to 1.17
≥20	1.22	0.91 to 1.63	1.21	0.81, 1.80	0.93	0.38 to 2.27
Follow-up for RCT						
End 3 years after enrolment	1.33	1.02 to 1.73			0.75	0.58 to 0.98
Begin 3 years after enrolment	1.26	0.92 to 1.74			0.96	0.72 to 1.27
Previous use of HT						
No	1.33	1.03 to 1.72				
Yes	1.21	0.86 to 1.71				
Propensity scores					0.88	0.70 to 1.12
<0.25	1.18	0.86 to 1.63	0.80	0.41 to 1.55	0.91	0.59 to 1.42
0.25–0.75	1.34	1.02 to 1.76	1.30	0.95 to 1.79	0.88	0.70 to 1.12
>0.75	2.57	0.78 to 8.43	0.50	0.19 to 1.32	0.79	0.41 to 1.51

HT, hormonal therapy; OS, observational study; RCT, randomised controlled trial; WHI, Women's Health Initiative.

participants taking E-alone the HR in the RCT was 1.44, and for participants taking E+P the HR was 1.53 for intervention participants. Risk-adjustment sometimes made HRs closer to 1.00 as expected (eg, intervention participants in the RCT for E+P), sometimes had minimal effect on HRs, and sometimes made a non-significant HR significant (eg, participants not on HT in the diet dataset).

Table 5 MI HRs comparing participants in each of the three RCT datasets to WHI OS participants

Outcome	Dataset	Unadjusted HR	Unadjusted χ^2	Adjusted† HR	Adjusted† χ^2
Patients not on HT (N=78 069)					
	RCT E+P	0.97	0.15	1.20	6.14
	RCT E alone	1.43***	24.98	1.37***	18.69
	RCT diet	1.01	0.07	1.14**	7.08
Patients on E+P (N=35 021)					
	RCT E+P	2.43***	97.23	1.53***	19.08
	RCT diet	1.29*	5.97	1.37**	8.86
Patients on E-only (N=38 672)					
	RCT E only	1.89***	58.17	1.44***	17.23
	RCT diet	0.99	0.00	1.04	0.22

*p<0.05.
**p<0.01.
***p<0.0001.
†Covariables used for the adjustment are described in the text.
E+P, oestrogen plus progestin; HT, hormone therapy; MI, myocardial infarction; OS, observational study; RCT, randomised controlled trial; WHI, Women's Health Initiative.

DISCUSSION

The WHI data analysed contained information on more than 800 possible confounders including information that made it possible to accurately predict HT use. It also contained information on factors that might have influenced response to HT. Some of these factors were related to the timing hypothesis (eg, age, time since menopause, previous HT use, beginning HT after baseline), and some were identified empirically (eg, blood pressure, previous coronary revascularisation and private medical insurance). Since OS and RCT participants differed with respect to these factors, these factors could have conceivably contributed to differences between the OSs and the RCTs. However, after taking into account all of these confounding factors and stratifying on factors that may have influenced the response to HT, OS and RCT differences remained.

The WHI data also contained information from four different studies, and the participants in these studies had different outcomes. After stratifying participants with respect to the type of HT and taking into account the information available in the WHI, we could not eliminate the outcome differences from the four studies.

The above results suggest that there were important risk factors not captured by the WHI that contributed to confounding. Since the WHI dataset is unusually comprehensive, it is likely that most OSs do not capture information on these risk factors. Without including information on potentially important confounders OSs cannot give reliably valid results.

Comparison to previous studies

OSs prior to WHI suggested a 30–50% reduction in coronary heart disease incidence among women using HT.[6–8] There was a smaller benefit shown in the analyses of the observational data in the present study: a 17% reduction in the OS and a 25% reduction in the RCT for diet.

After the WHI results were published, six studies of the association between HT and stroke or MI compared RCT results from the WHI with observational study results: three of these studies used observational data from the WHI[13 15 16] and three used observational data from the Nurses' Health Study.[9 10 14] Two of the WHI studies found, after controlling for time on HT and covariables, E+P HRs for MI did not significantly differ for the two study designs but HRs for stroke were higher in the RCT.

The goals and analytic methods of the present study differ substantially from previous studies using WHI data. The lead author believed that the extensive WHI data would be sufficient to give reliably valid results and extraordinary efforts were made to confirm this hypothesis. These efforts included an assessment of more than 800 risk factors as potential confounders and evaluating all marginally significant or previously suggested factors as potential effect modifiers. Even when the OS and RCT results were not the same, it was possible that the OS results were still valid. As a more definitive test of the adequacy of the WHI data we tried to eliminate differences in risk-adjusted outcomes from different datasets, which few if any other studies have attempted.

The present study differed from previous WHI studies in the following ways: (1) it included participants with and without a uterus, which made it possible to assess the effect of HT preparation. (2) It included participants in the diet RCT, which made it possible to compare risk-adjusted outcomes for two RCT and two OS datasets. (3) It evaluated more than 800 possible risk factors including those often suspected to cause confounding such as socioeconomic status, health behaviours, life style, stress and psychological characteristics. (4) It screened numerous participant characteristics for possible modulating effects on the association between HT and outcomes. (5) It analysed the risk for OS participants who began taking HT after enrolment. (6) It compared participants on the same treatment in different datasets and demonstrated that adjusting for WHI variables does not necessarily eliminate risk differences between datasets.

One of the WHI studies previously evaluated the timing hypothesis and did not find effects of prior HT use or menopause within 5 years.[16] Another analysis of WHI data has been often cited as supporting the timing hypothesis.[17] Although we tried to define coronary heart disease and years since menopause to get the same results, we could not. This suggests that the trends in the previous analysis were not robust to changing definitions.

A WHI study also found, as we did in the present study that the MI HR for E+P was greatest in the early years of treatment. This could explain OS and RCT differences because most OS participants taking HT at baseline began HT several years prior to baseline. However, some analyses in the present study did not support this explanation: (1) the RCT did not find that the effect of E-alone on MI changed over time; (2) none of the datasets found that the effect of any HT on stroke changed over time; (3) WHI OS participants who began HT after baseline had low MI risk and (4) prior HT exposure did not reduce the association between HT and cardiovascular disease.

Results from the OS performed by the Nurses' Health Study differed from our analysis of the WHI OS in important respects. One was that there was no protective association of HT and CHD for women over the age of 60.[9] (Other studies have also suggested that HT is less protective for older women.[11 12]) A second was that there was increased risk for new initiators of HT during the first 2 years after initiation and the risk increased 10 years after menopause.[14] Based on these findings the researchers in the Nurses' Health Study hypothesised that the OS results might be influenced by timing of HT initiation in relation to menopause onset or age and by length of follow-up. A third result that differed from ours was that HT significantly increased the risk of stroke.[10] Since this later result was similar to the WHI RCTs and the previous results might have explained differences between OSs and RCTs, the Nurses' Health Study suggested that OSs of HT could get the same results as RCTs.

The disagreements between our results and the results of the Nurses' Study do not show that the analyses or interpretation in either study are necessarily incorrect. The disagreements do demonstrate, however, the difficulty of getting valid results from OSs.

In addition to OSs of the Nurses' Health Study that give results similar to RCTs there is also an RCT that found oestradiol had an extraordinary protective effect on cardiovascular disease, which is consistent with the weaker protective effect of a different oestrogen preparation in the WHI OS.[12]

A previously published analysis of the WHI data shows that WHI risk factors cannot eliminate the association of adherence to placebo with MI, stroke or breast cancer.[18] Since the effect of adherence to placebo is probably a marker of unmeasured confounders, that study supports the implication of the present study that WHI risk factors are inadequate to eliminate unmeasured confounders.

Limitations

This study provided strong evidence that the WHI did not collect information on important risk factors related to MI or stroke. Although the WHI is unusually comprehensive, other datasets may provide information about these risk factors or about the risk factors that could

cause confounding for the outcomes they assessed. It is also possible that the WHI did collect the necessary information on the confounding factors, but the analytic methods used here were inadequate to take advantage of this information. However, the concerns raised by this study are still valid because both the dataset and the analytic methods used were much more comprehensive than is practical for almost all OSs.

Conclusion and future directions

We did not find that the comprehensive data provided by the WHI were adequate to overcome problems often attributed to OSs. The findings do not imply that most OSs are invalid. They do suggest, however, that given the current methodology, even very good OS datasets may not be adequate to give reliably valid results.

Owing to the key role that OSs are likely to play in studies of comparative effectiveness, it is critical to find ways to make OSs more valid. Although there has been some research on OS methodology,[14] more is required. There should be investigations to learn why some OSs agree with RCTs and others do not. More specific research goals include the following: (1) identify criteria for treatments unlikely to have confounding problems (eg, when there is little patient input to treatment, and one treatment is not preferred for higher risk patients), (2) find new risk factors that better adjust for patient behaviours that affect outcomes (eg, factors related to choosing or adhering to treatment) and (3) develop methods for assessment of confounding after data collection (eg, finding good markers for important unmeasured confounding factors). Without better OS methodology there will be underuse or misuse of OSs for comparative effectiveness research.

Author affiliations
[1]Huntsman Cancer Institute at the University of Utah, St Louis, Missouri, USA
[2]Health Services Research, University of Utah College of Medicine, Salt Lake City, Utah, USA
[3]Department of Epidemiology, College of Public Health, University of Iowa, Iowa City, Iowa, USA
[4]Department of Medicine, George Washington University School of Medicine, District of Columbia, DC, USA

Acknowledgements The Women's Health Initiative Study (WHI) is conducted and supported by the NHLBI in collaboration with the WHI Investigators. This manuscript was prepared using a limited access dataset obtained by the NHLBI and does not necessarily reflect the opinions or views of the WHI or the NHLBI. The research was supported in part by the Huntsman Cancer Foundation and the Beaumont Foundation.

Contributors AH supervised the study and prepared the manuscript. AH, TH, RW and JP participated in the conception and design, interpretation of data, revising the article and final approval of the version submitted.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement A de-identified dataset that contains all of the information used for the current study can be obtained by applying to the NHLBI.

REFERENCES

1. Feinstein AR. Epidemiologic analyses of causation: the unlearned scientific lessons of randomized trials. *J Clin Epidemiol* 1989;42:481–9. discussion 99–502.
2. Moses LE. Measuring effects without randomized trials? Options, problems, challenges. *Med Care* 1995;33(4 Suppl):AS8–14.
3. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;342:1878–86.
4. Investigators WGftWsHI. Risks and benefits of estrogen plus progestin in healthy postmenopausal women. *JAMA* 2002;288:321–33.
5. Anderson GL, Limacher M, Assaf AR, *et al*. Effects of conjugated equine estrogen in postmenopausal women with hysterectomy: the Women's Health Initiative randomized controlled trial. *JAMA* 2004;291:1701–12.
6. Barrett-Connor E, Grady D. Hormone replacement therapy, heart disease, and other considerations. *Annu Rev Public Health* 1998;19:55–72
7. Grady D, Rubin SM, Petitti DB, *et al*. Hormone therapy to prevent disease and prolong life in postmenopausal women. *Ann Intern Med* 1992;117:1016–37.
8. Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. *Prev Med* 1991;20:47–63
9. Grodstein F, Manson JE, Stampfer MJ. Hormone therapy and coronary heart disease: the role of time since menopause and age at hormone initiation. *J Womens Health (Larchmt)* 2006;15:35–44
10. Grodstein F, Manson JE, Stampfer MJ, *et al*. Postmenopausal hormone therapy and stroke: role of time since menopause and age at initiation of hormone therapy. *Arch Intern Med* 2008;168:861–6.
11. Salpeter SR, Cheng J, Thabane L, *et al*. Bayesian meta-analysis of hormone therapy and mortality in younger postmenopausal women. *Am J Med* 2009;122:1016–22. e1.
12. Schierbeck LL, Rejnmark L, Tofteng CL, *et al*. Effect of hormone replacement therapy on cardiovascular events in recently postmenopausal women: randomised trial. *BMJ* 2012;345:e6409.
13. Prentice RL, Langer RD, Stefanick ML, *et al*. Combined analysis of Women's Health Initiative observational and clinical trial data on postmenopausal hormone treatment and cardiovascular disease. *Am J Epidemiol* 2006;163:589–99.
14. Hernan MA, Alonso A, Logan R, *et al*. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* 2008;19:766–79.
15. Prentice RL, Langer R, Stefanick ML, *et al*. Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women's Health Initiative clinical trial. *Am J Epidemiol* 2005;162:404–14.
16. Prentice RL, Manson JE, Langer RD, *et al*. Benefits and risks of postmenopausal hormone therapy when it is initiated soon after menopause. *Am J Epidemiol* 2009;170:12–23.
17. Rossouw JE, Prentice RL, Manson JE, *et al*. Postmenopausal hormone therapy and risk of cardiovascular disease by age and years since menopause. *JAMA* 2007;297:1465–77.
18. Hartz A, He T. Why is greater medication adherence associated with better outcomes. *Emerg Themes Epidemiol* 2013. Published Online 2 Feb 2013. doi:10.1186/1742-7622-10-1