Randomized trials versus observational studies

The case of postmenopausal hormone therapy and heart disease

Miguel Hernán
Harvard School of Public Health
www.hsph.harvard.edu/causal

Joint work with James Robins and Luis García Rodríguez
The issue

- Observational studies found a lower risk of coronary heart disease (CHD) in users of postmenopausal hormone therapy
  - Nurses’ Health Study (NHS), General Practice Research Database (GPRD), 1980s and 1990s
  - Hazard ratios: 0.5-0.7

- A randomized trial found a greater risk of CHD in users
  - Women’s Health Initiative (WHI), 2002
  - Hazard ratio: 1.24
Chain reaction

- There is a clear discrepancy
- Randomized trials are the gold standard for causal inference
- Observational studies got it wrong
- Can observational studies ever be trusted again?
  - The end of observational epidemiology?
- Should we fund observational studies?
Our question of interest

- Do observational studies and randomized trials of hormone therapy and CHD provide the same answer if they are designed and analyzed in comparable ways?

- Before addressing this methodological issue, 3 clarifications
Clarification 1:
No discrepancy for all outcomes

- Results from WHI and observational studies are consistent for many other outcomes
  - Stroke
  - Venous thromboembolism
  - Breast cancer
  - Colon cancer
  - Hip fracture
  - etc...
Clarification 2: Different age distribution

- WHI participants are older on average than participants in the observational studies
- This may be important if the effect of hormones varies depending on the time between menopause and treatment initiation
  - e.g., hormones may increase the risk of CHD mostly in women at a more advanced stage of atherosclerosis
Clarification 3:
Two WHI randomized trials

- Both interrupted early
- First WHI trial
  - Manson et al. NEJM 2003
  - Estrogens plus progestin or placebo
  - HR of CHD: 1.24 (1.00, 1.54) or (0.97, 1.60)
- Second WHI trial
  - WHI Steering Committee. JAMA 2004
  - Estrogens (unopposed by progestin) or placebo
  - HR of CHD: 0.91 (0.75, 1.12)
Methodological explanations of the CHD discrepancy

Limitations of the observational studies
1. Lack of comparability between women who initiated and did not initiate hormone therapy
2. Lack of comparability between women who continued and discontinued hormone therapy
3. Uncertain time of therapy initiation
1. Lack of comparability between initiators and non initiators

- Can explain observational results if initiators had, say, a healthier lifestyle than non initiators
  - Confounding for treatment initiation

- In WHI, no confounding because treatment initiation was randomly assigned
2. Lack of comparability between those who did and did not continue

- Can explain observational results if women who stayed on therapy were, say, more health-conscious than the others
  - Confounding for treatment discontinuation

- Can arise in WHI because treatment discontinuation was not randomized, but cannot explain WHI results because analysis followed intention-to-treat principle
3. Uncertain time of therapy initiation

- NHS data collected every 2 years
  - women assigned to the hormone use group reported in the questionnaire at the onset of the 2-year interval
  - initiators systematically misclassified as non-users until the next questionnaire
  - if initiation causes a short-term increase in risk, then estimate downwardly biased

- In the WHI (randomized) and the GPRD (observational), no such uncertainty
  - Time of hormone therapy initiation is known
Our proposed strategy (I)

- Re-analyze the observational studies without limitations 2 and 3
- Then compare new estimates with that of the randomized trial
Our proposed strategy (II)

- Remove limitation 2 by re-analyzing the data under an intention-to-treat principle
  - Need to conceptualize the observational study as a sequence of “randomized” trials
  - Randomization probabilities are unknown but can be estimated from the data

- First re-analyze GPRD (no limitation 3), then re-analyze NHS (deal with limitation 3 by sensitivity analysis)
The GPRD

- Research-oriented database that covers over 3 million residents in the UK
- These individuals' general practitioners register healthcare and medical information about their patients in a standardized manner
  - demographic data, medical diagnoses, consultant and hospital referrals, prescriptions
  - Validation studies: 90% of information in the paper medical records and 95% of newly prescribed drugs are in the database
The GPRD trials
Eligibility criteria

- Eligibility criteria
  - women aged 50 years or more and with an intact uterus
  - Exclusion criteria: past diagnosis of cancer (except non melanoma skin cancer), cardiovascular disease, and cerebrovascular disease

- Similar to WHI criteria
The GPRD trials
Baseline and follow-up

- Follow-up
  - From **baseline** to the diagnosis of a CHD endpoint, death from causes other than CHD, loss to follow-up, or administrative end of follow-up, whichever came first

- What is baseline?
  - WHI: women followed from time of randomization
  - GPRD: women followed from January 1991
The GPRD trials
Treatment regimes

1) Initiation of use of oral estrogens plus progesterone at baseline
2) No hormone use at baseline

☐ Washout interval: non users in 1-yr period before baseline (additional eligibility criterion)
The GPRD trials
Intention to treat (ITT) principle

- Compare the risk of CHD between women who initiated and did not initiate hormone therapy at baseline
- Regardless of future hormone use during the follow-up
The GPRD Trials
Analytic approach

- Cox proportional hazards model
- Covariates:
  - Indicator for hormone therapy initiation
  - Age, calendar month, family history of CHD, high cholesterol, high blood pressure, diabetes, body mass index, smoking, alcohol intake, aspirin use, nonsteroidal antiinflammatory drug use, previous use of hormone therapy
- Conditional ITT hazard ratio
The GPRD trials

- There is nothing special about January 1991
- We can start our trial in February 1991, March 1991, ..., December 2001
  - Sequence of nested trials
- Or we can conduct all possible trials, pool the data across trials, and obtain an effect estimate with a narrower confidence interval
  - Need to adjust the variance
- Eligibility criteria applied at each trial baseline
The GPRD trials

- We started a separate GPRD trial at each month $m$
  - $m=0,1,...,131$ representing January 1991, February 1991,..., December 2001
- Each woman may participate in a maximum of 132 trials
- For each trial, follow-up started in month $m$ (baseline) and ended at diagnosis of a CHD endpoint, death, lost to follow-up, or administrative end of follow-up (eight years like in the WHI or December 2001), whichever came first
Analytic approach
(Nested) Cox model

\[ \lambda_T[t|G(m) = 1, A(m), \bar{L}(m)] = \lambda_0[t][\alpha A(m) + \theta'_1 \bar{L}(m)] \]

- Notation
  - \( T \): CHD-free survival time
  - \( G(m) \): indicator for eligibility at \( m \)
  - \( L(m) \): covariates measured before \( m \)

- PMLE, robust variance
- Conditional ITT hazard ratio: \( \exp(\alpha) \)
- Similar results using doubly-robust estimators from nested structural AFT model that incorporates propensity score
Results
Women eligible for trials

- 99,072 women contributed to trials
  - 1,889 cases
  - 606 dead

Pooling over trials
- On average, each woman participated in 60.5 trials
- 5,997,824 participants
- 10,566 initiators
Example: Trials 25-50

<table>
<thead>
<tr>
<th>Trial</th>
<th>Month</th>
<th>Participants</th>
<th>CHD events</th>
<th>Initiators</th>
<th>CHD events in initiators</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>January 1993</td>
<td>68,026</td>
<td>1134</td>
<td>218</td>
<td>1</td>
</tr>
<tr>
<td>26</td>
<td>February 1993</td>
<td>67,774</td>
<td>1112</td>
<td>193</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>March 1993</td>
<td>67,669</td>
<td>1085</td>
<td>239</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>April 1993</td>
<td>67,338</td>
<td>1060</td>
<td>201</td>
<td>1</td>
</tr>
<tr>
<td>29</td>
<td>May 1993</td>
<td>66,972</td>
<td>1030</td>
<td>200</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>June 1993</td>
<td>66,893</td>
<td>1009</td>
<td>170</td>
<td>1</td>
</tr>
<tr>
<td>31</td>
<td>July 1993</td>
<td>66,720</td>
<td>985</td>
<td>168</td>
<td>0</td>
</tr>
<tr>
<td>32</td>
<td>August 1993</td>
<td>66,655</td>
<td>966</td>
<td>192</td>
<td>0</td>
</tr>
<tr>
<td>33</td>
<td>September 1993</td>
<td>66,354</td>
<td>947</td>
<td>134</td>
<td>1</td>
</tr>
<tr>
<td>34</td>
<td>October 1993</td>
<td>66,301</td>
<td>928</td>
<td>132</td>
<td>0</td>
</tr>
<tr>
<td>35</td>
<td>November 1993</td>
<td>66,165</td>
<td>908</td>
<td>155</td>
<td>1</td>
</tr>
<tr>
<td>36</td>
<td>December 1993</td>
<td>65,983</td>
<td>884</td>
<td>98</td>
<td>0</td>
</tr>
<tr>
<td>37</td>
<td>January 1994</td>
<td>69,729</td>
<td>871</td>
<td>149</td>
<td>2</td>
</tr>
<tr>
<td>38</td>
<td>February 1994</td>
<td>69,592</td>
<td>858</td>
<td>185</td>
<td>2</td>
</tr>
<tr>
<td>39</td>
<td>March 1994</td>
<td>69,262</td>
<td>833</td>
<td>196</td>
<td>3</td>
</tr>
<tr>
<td>40</td>
<td>April 1994</td>
<td>69,019</td>
<td>813</td>
<td>168</td>
<td>0</td>
</tr>
<tr>
<td>41</td>
<td>May 1994</td>
<td>68,919</td>
<td>801</td>
<td>141</td>
<td>0</td>
</tr>
<tr>
<td>42</td>
<td>June 1994</td>
<td>68,442</td>
<td>785</td>
<td>146</td>
<td>1</td>
</tr>
<tr>
<td>43</td>
<td>July 1994</td>
<td>68,245</td>
<td>751</td>
<td>135</td>
<td>0</td>
</tr>
<tr>
<td>44</td>
<td>August 1994</td>
<td>68,053</td>
<td>736</td>
<td>158</td>
<td>0</td>
</tr>
<tr>
<td>45</td>
<td>September 1994</td>
<td>67,769</td>
<td>718</td>
<td>137</td>
<td>2</td>
</tr>
<tr>
<td>46</td>
<td>October 1994</td>
<td>67,681</td>
<td>689</td>
<td>135</td>
<td>1</td>
</tr>
<tr>
<td>47</td>
<td>November 1994</td>
<td>67,413</td>
<td>661</td>
<td>145</td>
<td>1</td>
</tr>
<tr>
<td>48</td>
<td>December 1994</td>
<td>67,151</td>
<td>648</td>
<td>97</td>
<td>0</td>
</tr>
<tr>
<td>49</td>
<td>January 1995</td>
<td>69,901</td>
<td>626</td>
<td>178</td>
<td>1</td>
</tr>
<tr>
<td>50</td>
<td>February 1995</td>
<td>69,500</td>
<td>618</td>
<td>146</td>
<td>1</td>
</tr>
</tbody>
</table>
Results

CHD hazard ratios and 95\% CI

<table>
<thead>
<tr>
<th>Years of follow-up</th>
<th>Initiators versus noninitiators</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model (1)</td>
</tr>
<tr>
<td></td>
<td>ITT</td>
</tr>
<tr>
<td>0–2</td>
<td>1.20 (0.84, 1.72)</td>
</tr>
<tr>
<td>0–5</td>
<td>0.99 (0.76, 1.28)</td>
</tr>
<tr>
<td>0–8</td>
<td>0.95 (0.75, 1.20)</td>
</tr>
<tr>
<td>All</td>
<td>0.92 (0.73, 1.17)</td>
</tr>
</tbody>
</table>

Closer to WHI estimates

1.68 first 2 years, 1.24 overall
Discussion: Assumption of no unmeasured confounding

- Adjustment for baseline variables had little effect on the estimates.
- Unadjusted HR 0.85 only moderately less than the fully-adjusted estimate 0.92.
  - The magnitude of confounding due to unmeasured variables would have to exceed the confounding due to measured variables to explain the full GPRD-WHI discrepancy.
  - A downward bias of 0.1-0.2 in our hazard ratio estimate is plausible.
Discussion

Sampling variability

- Major problem
- Overall ITT hazard ratios from the GPRD and the WHI trials were estimated with similar low precision
  - width of the 95% CIs on the log scale: about 0.46 in WHI and 0.47 in GPRD
- This relatively low precision precludes drawing strong conclusions from either study
  - Overlapping confidence intervals
Conclusions regarding the discrepancy WHI/GPRD

- Under our analytic approach, smaller difference between randomized and observational estimates
- Consistent with
  - small amount of unmeasured confounding
  - random variability
- Had the GPRD analyzed under our approach, WHI results would not have been that surprising
  - This approach is a particular case of Robins’ g-estimation
Methodological conclusions

- A direct comparison between the estimates of observational studies and randomized trials can be misleading
  - Randomized trials analyzed under ITT principle
  - Observational studies analyzed using the ‘as treated’ principle

- Fair comparison requires a comparable analytic approach like the one described here
  - NHS is next
But there is something else

- ITT estimates comparable between studies only if similar “noncompliance” rates
- Is that true?
- At 6 years
  - Therapy discontinuation in initiators: 42% WHI, 79% GPRD
  - Therapy initiation in non initiators: 11% WHI, 13% GPRD
Large differences in noncompliance

- Reasonable to compare their ITT effect estimates?
- In addition, no guarantee that the GPRD and WHI “noncompliers” were comparable
- Problem for comparisons between
  - Observational and randomized studies
  - Randomized and randomized studies
- Need to go beyond ITT effects: estimate effect of continuous treatment that would be observed under full compliance
Estimation of the effect of continuous treatment (I)

- Artificial censoring plus inverse probability weighting
- In each trial $m$, we censored women when they discontinued their baseline treatment
- Because this censoring is potentially informative and may lead to selection bias, a woman uncensored at $k$ was upweighted by the inverse of their estimated probability of remaining uncensored from month $m$ through month $k$
Estimation of the effect of continuous treatment (II)

- We estimated this probability separately in initiators and non initiators by fitting the logistic models

\[ \text{logit } \Pr [A(j) = a \mid G(m) = 1, \quad A(j - 1) = a, A(m) = a, \bar{L}(j), T > j] = \theta_{a0} + \theta'_{a1} \bar{L}(j) \quad \text{for} \quad j > m, \]
## Results

<table>
<thead>
<tr>
<th>Years of follow-up</th>
<th>Initiators versus noninitiators</th>
<th>Continuous versus never users</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model (1) ITT</td>
<td>Model (1) IPW</td>
</tr>
<tr>
<td>0–2</td>
<td>1.20 (0.84, 1.72)</td>
<td>1.33 (0.79, 2.22)</td>
</tr>
<tr>
<td>0–5</td>
<td>0.99 (0.76, 1.28)</td>
<td>0.83 (0.52, 1.32)</td>
</tr>
<tr>
<td>0–8</td>
<td>0.95 (0.75, 1.20)</td>
<td>0.95 (0.60, 1.51)</td>
</tr>
<tr>
<td>All</td>
<td>0.92 (0.73, 1.17)</td>
<td>0.87 (0.55, 1.39)</td>
</tr>
</tbody>
</table>
Next step

- Since method can be applied to
  - Randomized trials
  - Observational studies

- Then, ideally, this adjustment for noncompliance would be applied to the WHI data as well