Methodological issues in observational studies of comparative effectiveness

Choice of comparators, target population, and hypothesis formulation

Robert J Glynn
Divisions of Pharmacoepidemiology & Pharmacoeconomics
Brigham & Women’s Hospital, Boston, MA
Methodological issues in observational studies of comparative effectiveness

Future talks

March: Martin Kulldorff
Sequential analysis applied to comparative effectiveness

April: Joshua Gagne
Comparative effectiveness of newly marketed drugs

May: Jessica Myers
Integrating effectiveness information across outcomes
Disclosure

Supported by grant AG018833 from the US National Institute on Aging

I have worked on grants from AstraZeneca and Novartis to the Brigham & Women’s Hospital for the design, monitoring and analysis of clinical trials; and signed a consulting agreement with Merck to give a grand rounds talk
Outline

1. Challenges to active comparators
2. Personal experience with non-user referents
3. Likely differential surveillance in non-users
5. Healthy starter/sick stopper bias
6. Time scale in safety/effectiveness studies
7. Propensity score trimming and focusing the target population
8. Composite endpoints and competing risks
Comparative Effectiveness Research (CER): Definition and Aims

The generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care.*

Aim: to improve decisions that affect medical care at the levels of both policy and the individual.†

†HC Sox, SN Goodman. Annu Rev Pub Health 2012
CER: Key elements

(a) Head-to-head comparisons of active treatments
(b) Study populations typical of day-to-day clinical practice
(c) Focus on evidence to inform care tailored to the characteristics of individual patients*

*HC Sox, SN Goodman. Annu Rev Pub Health 2012
Example: oral diabetes drugs and cancer risk

Diabetes is associated with increased risk of several common cancers*

Association may be partly due to shared risk factors such as aging, obesity, diet, and physical inactivity*

Evidence on relationships of specific antidiabetic drugs with cancer risk is limited, and possibly confounded

Early evidence suggests that metformin is associated with reduced risk of cancer

Important time scales

Age in years:

risk factor for DM and cancer
often the best scale for risk set construction (EL Korn et al. Am J Epidemiol 1997)

Time since DM Dx:

related to cancer risk and surveillance
likely measured with differential error

Time since Rx initiation:

natural scale for Rx evaluation
problematic for switchers and those with gaps
Errors in assessing duration of DM

Onset date is left-censored for subjects who enter the database with DM

DM is often undiagnosed (about 27% of cases currently; formerly 50% in the US)

DM is often not recorded among discharge diagnoses, and is less often recorded in those near death

Drug treatment for DM is less frequent in frail people

Selective recording and treatment of DM

Time since Rx initiation

Often most relevant in pharmacoepidemiology; mimics time after randomization in a trial

Risk sets make comparisons (with active comparators) at equivalent treatment durations

Cumulative dose consumed can vary for given time because of gaps and differences in daily dose

Variant scale restricts to those with at least one (or two) refills and starts follow-up at that time

For second line therapy, duration of prior drug treatment an important covariate
Time-related biases in studies of DM and cancer

Attention to time scales can prevent time-related biases, which may have distorted treatment effects in several studies finding a protective effect of metformin*

Immortal time bias: introduced with time-fixed cohort analyses that misclassify unexposed time as exposed

Time-window bias: differential exposure opportunity time windows between subjects

Time-lag bias: can occur in comparisons of treatments given at different stages of disease

S Suissa, L Azoulay. Metformin and the risk of cancer: time-related biases in observational studies. Diab Care 2012
Alternative comparison groups/time scales

Non-diabetic individuals followed from a specific time point

Non-diabetic initiators of another drug

Untreated diabetics

Diabetic individuals followed from initiation of a comparison drug
Limitations of non-diabetic comparators

DM is an independent risk factor for cancer

Unmeasured lifestyle confounders (diet, exercise, SES, BMI) are unbalanced

Unmeasured biomarker confounders (fasting glucose, glycated hemoglobin)*

Likely differential surveillance, barriers to diagnosis/treatment

*Emerging risk factors collaboration. Diabetes mellitus, fasting glucose, and risk of cause-specific death. NEJM 2011
Limitations of non-DM drugs as comparators

Use of active comparators leads to risk sets based on time-on-therapy in all study subjects.

Can help account for barriers to persistence and balance surveillance.

Different indications to initiation and barriers to persistence can limit use of non-diabetic drugs as control exposures.

Also, levels of DM risk factors and attention to their surveillance likely differ in non-diabetics.
Limitations of diabetic non-user comparators

Diet-controlled diabetes typically less severe

Non-user referents lack treatment initiation and termination dates

Can have decreased surveillance for comorbidities relative to medication users

Often missing data on barriers to initiation and persistence

Problematic to control for duration of treatment
Distinguish placebo controls and non-user referents

Placebo controls are often natural, appropriate referents for evaluation of safety and efficacy in randomized trials.

Non-user referents in observational studies differ in important ways:

Not clearly eligible or willing to initiate therapy

No clear date of initiation

No comparable assessment of treatment duration
Challenges to active comparators

Other DM drugs may influence cancer risk; e.g. possible protective effects of metformin

Or be initiated after different durations of DM, according to guidelines*

Or in patients that differ on unmeasured cancer risk factors (e.g. BMI)

*DM Nathan et al. Medical management of hyperglycemia in type 2 diabetes: a consensus algorithm for the initiation and adjustment of therapy. Diabetes Care 2009
Advantages of DM drugs as comparators

In spite of guidelines, uncertainty about treatments leads to substantial variability in actual practice.

Comparisons among active, within-class drugs address comparative safety & efficacy questions.

Comparisons restricted to treated patients with diabetes offer a better chance for confounder control, especially in claims data.
Three paradoxes with non-user referents

1. Among hospitalized, older people enrolled in state-sponsored drug benefits programs, diabetes diagnosis and treatment were associated with enhanced survival vs no dx or rx.

2. Across 20 commonly used classes of drugs, users vs non-users of several classes had markedly reduced death rates, of a magnitude inconsistent with randomized evidence.

3. Focused analysis (e.g. propensity score matching) did little to reduce the magnitude of the incredible reduction in the hazard of death in users vs non-users of lipid-lowering drugs.
Observational data on mortality with non-user referent

Glynn et al. Paradoxical relations of drug treatment with mortality in older persons. Epidemiology 2001
Observational data on mortality with non-user referent

Use of lipid-lowering drugs and the hazard of death in the community sample

<table>
<thead>
<tr>
<th></th>
<th>Hazard ratio</th>
<th>95% CI</th>
<th>Change in hazard ratio from base estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unmatched design: proportional hazards analysis with 17,524 deaths among 106,838 beneficiaries</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted for age and sex</td>
<td>0.44</td>
<td>0.40–0.49</td>
<td>Base estimate</td>
</tr>
<tr>
<td>Further adjusted for demographics and comorbidity indices</td>
<td>0.49</td>
<td>0.44–0.54</td>
<td>11%</td>
</tr>
<tr>
<td>Further adjusted for propensity quintile</td>
<td>0.54</td>
<td>0.48–0.60</td>
<td>23%</td>
</tr>
<tr>
<td>Further adjusted for use of other drugs and comorbidity</td>
<td>0.61</td>
<td>0.55–0.68</td>
<td>40%</td>
</tr>
<tr>
<td>Matched design: Proportional hazards analysis with 966 deaths in 6,466 users of lipotropics and propensity-matched nonusers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted for age and sex</td>
<td>0.60</td>
<td>0.53–0.69</td>
<td></td>
</tr>
<tr>
<td>Further adjusted for use of other drugs and comorbidity</td>
<td>0.60</td>
<td>0.52–0.68</td>
<td></td>
</tr>
</tbody>
</table>

Based on McNemar’s matched analysis

*Abbreviations*: CI, confidence interval.

Glynn et al. Selective prescribing led to overestimation of the benefits of lipid lowering drugs. J Clin Epidemiol 2006
Variable surveillance in effectiveness/safety studies: exposure assessment

Initiation of treatments is generally measured well in administrative databases (although whether subjects actually take medications is unknown)

Duration of exposure indexed by prescription refills

Assessment of exposure duration is problematic with coverage gaps (e.g. nursing home stays)

Non-user referents lack initiation dates or comparable assessment of duration of exposure
Variable surveillance in effectiveness/safety studies: outcome assessment

Exposed subjects are often followed more intensively

Especially intensive for anticipated outcomes such that these may not be evaluable in observational designs*

Even for unanticipated effects, prescribers may follow treated patients more intensively than non-user referents

Randomized trials often block on center as a way to balance outcome surveillance

Variable surveillance in effectiveness/safety studies: covariate assessment

Treatment initiators generally have more intensive surveillance than non-initiators

Database studies usually assume comorbid conditions are absent unless expressly noted

Comorbidities related to treatment decisions may be particularly scrutinized

Little prior study of whether covariate scrutiny is jointly related to treatment and disease risk
Design strategies to balance surveillance in effectiveness/safety studies

Often studies restrict populations to subjects with demonstrated eligibility and those who meet some system use criteria (e.g. history of filled prescriptions through the system)

Can also reduce surveillance variability by restriction to subjects with documented risk levels for the outcome; e.g. prior MI or diagnosed ESRD

Even with such restrictions, treatment initiators will likely have greater surveillance than non-initiators
## Differential covariate assessment in non-users

### Table 1: Baseline characteristics of study population of new users and non-users of statins. Values are numbers (percentages) unless stated otherwise

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>New users (n=225,922)</th>
<th>Non-users (n=1,778,770)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women</td>
<td>104,774 (46.4)</td>
<td>909,423 (51.1)</td>
</tr>
<tr>
<td>Men</td>
<td>121,148 (53.6)</td>
<td>869,347 (48.9)</td>
</tr>
<tr>
<td>Ethnicity:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recorded</td>
<td>121,355 (53.7)</td>
<td>569,466 (32.0)</td>
</tr>
<tr>
<td>White or not stated</td>
<td>215,077 (95.2)</td>
<td>1,699,991 (95.6)</td>
</tr>
<tr>
<td>Indian</td>
<td>2,861 (1.3)</td>
<td>13,398 (0.8)</td>
</tr>
<tr>
<td>Pakistani</td>
<td>16,558 (0.7)</td>
<td>7,562 (0.4)</td>
</tr>
<tr>
<td>Bangladeshi</td>
<td>679 (0.3)</td>
<td>3,226 (0.2)</td>
</tr>
<tr>
<td>Other Asian</td>
<td>759 (0.3)</td>
<td>7,321 (0.4)</td>
</tr>
<tr>
<td>Caribbean</td>
<td>17,888 (0.8)</td>
<td>9,853 (0.6)</td>
</tr>
<tr>
<td>Black African</td>
<td>8,344 (0.4)</td>
<td>15,358 (0.9)</td>
</tr>
<tr>
<td>Chinese</td>
<td>316 (0.1)</td>
<td>4,015 (0.2)</td>
</tr>
<tr>
<td>Other ethnic group</td>
<td>19,500 (0.9)</td>
<td>18,046 (1.0)</td>
</tr>
</tbody>
</table>
### Differential covariate assessment in non-users

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value 1</th>
<th>Value 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD) age (years)</td>
<td>57.2 (11.7)</td>
<td>44.4 (13.7)</td>
</tr>
<tr>
<td>Mean (SD) Townsend score</td>
<td>-0.5 (3.3)</td>
<td>-0.3 (3.4)</td>
</tr>
<tr>
<td>Mean (SD) systolic blood pressure (mm Hg)</td>
<td>141.1 (19.1)</td>
<td>129.9 (19.1)</td>
</tr>
<tr>
<td>Liver function test:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recorded at baseline or before statins</td>
<td>131 354 (58.1)</td>
<td>162 207 (51.1)</td>
</tr>
<tr>
<td>Recorded at follow-up</td>
<td>193 586 (85.7)</td>
<td>594 750 (33.4)</td>
</tr>
<tr>
<td>Creatine kinase concentration:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recorded at baseline or before statins</td>
<td>15 724 (7.0)</td>
<td>8642 (0.5)</td>
</tr>
<tr>
<td>Recorded at follow-up</td>
<td>62 706 (27.8)</td>
<td>43 333 (2.4)</td>
</tr>
<tr>
<td>Body mass index recorded</td>
<td>207 644 (91.9)</td>
<td>1 341 863 (75.4)</td>
</tr>
<tr>
<td>Mean (SD) body mass index</td>
<td>28.3 (4.9)</td>
<td>26 (4.6)</td>
</tr>
<tr>
<td>Smoking status recorded</td>
<td>224 982 (99.6)</td>
<td>1 615 527 (90.8)</td>
</tr>
<tr>
<td>Body mass index and smoking status recorded</td>
<td>207 494 (91.8)</td>
<td>1 330 320 (74.8)</td>
</tr>
</tbody>
</table>
Does comparability of information between exposed and unexposed matter?

Not necessarily!

Comparability-of-information principle dictates the importance of equivalent information in compared groups*

Rothman & Greenland† argue against the principle

Nondifferential measurement error does not guarantee that bias is toward the null

e.g. a case-control study using proxy respondents for cases might have greater bias with proxy responses for controls

Does confounder measurement error matter?

Investigators often focus on accurate exposure and outcome assessment, with a more sanguine view of errors in confounders*

Non-differential confounder misclassification biases toward the crude, can induce artificial heterogeneity of effects across confounder categories, but is somewhat limited in magnitude†

Differential confounder misclassification can bias in either direction and yield much greater heterogeneity in estimates across confounder categories.

Sensitivity analyses are needed to evaluate potential impact of inaccurate confounder assessment

Example with differential confounder assessment

Consider data on 6-month risk of death (D=0 vs 1) in lung cancer patients treated at the same CT hospitals in 2 time periods (P=0 vs 1), separated by a 13-yr interval.*

Of interest is whether risk is lower in the later period due to treatment advances.

P & D are measured without error

A potential confounder is TNM (tumor, nodes, and metastases) stage (S=1, 2, or 3) at baseline, likely measured differently at baseline due to advances in imaging technologies.

Errors in S are related to P and possibly also risk of D

Impact of confounder measurement error
Hypothetical study

Consider a prospective study of statin use \((S=0 \text{ vs } 1)\) and occurrence of venous thromboembolism \((VTE=0 \text{ or } 1)\)

Obesity \((O=0 \text{ or } 1)\) is strongly related to statin use and VTE risk

Available is a surrogate of \(O\), \(O^*\), measured with sensitivity \(P_r(O^*=1|O=1)\) \(S_e\), and specificity \(P_r(O^*=1|O=0)\) \(S_p\)

\(S_e\) and \(S_p\) might be non-differential, or differential, i.e. dependent on \(S\), \(VTE\), or both
### Impact of confounder measurement error

#### Hypothetical true data

<table>
<thead>
<tr>
<th></th>
<th>VTE=1</th>
<th>VTE=0</th>
<th>Total</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>O=0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S=1</td>
<td>10</td>
<td>490</td>
<td>500</td>
<td>.02</td>
</tr>
<tr>
<td>S=0</td>
<td>190</td>
<td>9310</td>
<td>9500</td>
<td>.02</td>
</tr>
<tr>
<td>O=1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S=1</td>
<td>60</td>
<td>940</td>
<td>1000</td>
<td>.06</td>
</tr>
<tr>
<td>S=0</td>
<td>240</td>
<td>3760</td>
<td>4000</td>
<td>.06</td>
</tr>
</tbody>
</table>

\[
\text{OR}_0 = 1.00 \quad \text{OR}_1 = 1.00 \quad \text{OR}_{\text{crude}} = 1.49 \quad \text{OR}_{\text{adj}} = 1.00
\]
Impact of confounder measurement error

Non-differential error: $O^*$ has $Se = 0.7; Sp = 0.8$

<table>
<thead>
<tr>
<th>$O^* = 0$</th>
<th>VTE = 1</th>
<th>VTE = 0</th>
<th>Total</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S = 1$</td>
<td>26</td>
<td>674</td>
<td>700</td>
<td>0.037</td>
</tr>
<tr>
<td>$S = 0$</td>
<td>224</td>
<td>8576</td>
<td>8800</td>
<td>0.025</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$O^* = 1$</th>
<th>VTE = 1</th>
<th>VTE = 0</th>
<th>Total</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S = 1$</td>
<td>44</td>
<td>756</td>
<td>800</td>
<td>0.055</td>
</tr>
<tr>
<td>$S = 0$</td>
<td>206</td>
<td>4494</td>
<td>4700</td>
<td>0.044</td>
</tr>
</tbody>
</table>

$OR_0 = 1.48$  $OR_1 = 1.27$  $OR_{\text{crude}} = 1.49$  $OR_{\text{adj}} = 1.34$
Differential confounder measurement error

\( S = 0: O^* \times Se = 0.6, Sp = 0.7; Se = 1: O^* \times Se = 0.8, Sp = 0.9 \)

<table>
<thead>
<tr>
<th></th>
<th>VTE=1</th>
<th>VTE=0</th>
<th>Total</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S=1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>O^*=0</strong></td>
<td>21</td>
<td>629</td>
<td>650</td>
<td>.032</td>
</tr>
<tr>
<td><strong>S=0</strong></td>
<td>229</td>
<td>8021</td>
<td>8250</td>
<td>.028</td>
</tr>
<tr>
<td><strong>S=1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>O^*=1</strong></td>
<td>49</td>
<td>801</td>
<td>850</td>
<td>.058</td>
</tr>
<tr>
<td><strong>S=0</strong></td>
<td>201</td>
<td>5049</td>
<td>5250</td>
<td>.038</td>
</tr>
</tbody>
</table>

\( OR_0 = 1.17 \quad OR_1 = 1.54 \quad OR_{\text{crude}} = 1.49 \quad OR_{\text{adj}} = 1.40 \)
Differential error in sensitivity only
\[ S = 0: O^* \text{ Se} = .5; S = 1: O^* \text{ Se} = .8 \]

<table>
<thead>
<tr>
<th></th>
<th>VTE=1</th>
<th>VTE=0</th>
<th>Total</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>*<em>O</em>=0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S=1</strong></td>
<td>22</td>
<td>678</td>
<td>700</td>
<td>.031</td>
</tr>
<tr>
<td><strong>S=0</strong></td>
<td>310</td>
<td>11190</td>
<td>11500</td>
<td>.027</td>
</tr>
<tr>
<td>*<em>O</em>=1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S=1</strong></td>
<td>48</td>
<td>752</td>
<td>800</td>
<td>.06</td>
</tr>
<tr>
<td><strong>S=0</strong></td>
<td>120</td>
<td>1880</td>
<td>2000</td>
<td>.06</td>
</tr>
</tbody>
</table>

\[ \text{OR}_0 = 1.17 \quad \text{OR}_1 = 1.00 \quad \text{OR}_{\text{crude}} = 1.49 \quad \text{OR}_{\text{adj}} = 1.06 \]
Differential error in sensitivity greater if $VTE = 0$
$Se$ for $(S,VTE): .5 (0,0) .6 (0,1), .8 (1,0), .9 (1,1)$

<table>
<thead>
<tr>
<th>$O^* = 0$</th>
<th>$VTE = 1$</th>
<th>$VTE = 0$</th>
<th>Total</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S = 1$</td>
<td>16</td>
<td>678</td>
<td>700</td>
<td>.023</td>
</tr>
<tr>
<td>$S = 0$</td>
<td>286</td>
<td>11190</td>
<td>11500</td>
<td>.025</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$O^* = 1$</th>
<th>$VTE = 1$</th>
<th>$VTE = 0$</th>
<th>Total</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S = 1$</td>
<td>54</td>
<td>752</td>
<td>800</td>
<td>.068</td>
</tr>
<tr>
<td>$S = 0$</td>
<td>144</td>
<td>1880</td>
<td>2000</td>
<td>.072</td>
</tr>
</tbody>
</table>

$OR_0 = 0.92 \quad OR_1 = 0.94 \quad OR_{crude} = 1.49 \quad OR_{adj} = 0.93$
## Time period (P), and death (D) by stage (S)

<table>
<thead>
<tr>
<th>StageA</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D=1</td>
<td>D=0</td>
<td>D=1</td>
</tr>
<tr>
<td>P=1</td>
<td>2</td>
<td>22</td>
<td>5</td>
</tr>
<tr>
<td>P=0</td>
<td>70</td>
<td>211</td>
<td>74</td>
</tr>
<tr>
<td>RR</td>
<td>.33</td>
<td></td>
<td>.65</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>StageB</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D=1</td>
<td>D=0</td>
<td>D=1</td>
</tr>
<tr>
<td>P=1</td>
<td>10</td>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>P=0</td>
<td>70</td>
<td>211</td>
<td>74</td>
</tr>
<tr>
<td>RR</td>
<td>.96</td>
<td>.74</td>
<td></td>
</tr>
</tbody>
</table>
Lung cancer stage, time period and risk of death

With stage classified using additional information in period 1 (StageA method), death risk was significantly lower in period 1 ($RR_{adj} 0.77; 95\% \text{ CI: } 0.64-0.93$), slightly stronger than $RR_{crude} 0.80$

With equivalent staging that ignored new imaging information (StageB method), each stratum-specific RR was attenuated, and protective effect was equivocal ($RR_{adj} 0.89; 95\% \text{ CI: } .74-1.07$)

Patients who migrated to a higher stage with additional imaging information were at higher risk than those remaining in their original stage but lower risk than others in their new stage

Feinstein et al dubbed the bias associated with StageA classification the “Will Rogers Effect”

Parallels between the healthy worker effect and biases in studies of preventive drug use

1. Healthy hire effect similar to selection of healthier individuals for preventive drug therapy: healthy starter effect

2. Healthy worker survivor effect similar to better adherence in healthier people: sick stopper effect

3. Inevitable aging effect applies in both situations
Ways to account for the healthy worker effect*

1. Restrict analysis to survivors of a fixed time period; e.g. 15 year survivors (removes healthy starter effect)

2. Lag the exposure to exclude those with recent exposure

3. Adjust for continued exposure as a confounder

4. But continued exposure may be on the causal pathway (both a confounder and an intermediary) suggesting marginal structural models (J Robins)

No consensus on the optimal approach

*HM Arrighi & I Hertz-Picciotto; Epidemiology 1994
The time scale for occupational and pharmacoepidemiologic studies

1. Time since hire relates to outcomes in two ways: the healthier workers stay longer in the workforce; yet their health inexorably declines with age

2. It is critical to control for time since hire*

3. Parallel considerations argue for the need to control for time since initiation and time on therapy in pharmacoepidemiology

*WD Flander et al. Epidemiology 1993
Selective use of preventive therapies

Complex factors influence use of preventive therapies

Healthy user bias arises from unmeasured determinants of use

Bias can be large and changed little by control for available covariates

Intractable to standard analytic strategies, not just in claims databases
The Healthy Starter Effect

Complex factors influence decisions to initiate preventive therapies

Providers may focus on a patient’s main or most symptomatic medical problem (DA Redelmeier et al N Engl J Med 1998)

Preventive strategies may be less beneficial with short life expectancy

Measurement and control for frailty is very difficult, so substantial confounding may remain with traditional approaches
The Sick Stopper Effect

Just as starting a preventive regimen is a sign of good health, sticking with a therapy is a sign of healthy behaviors and good health.

Determinants of stopping may differ from and be more difficult to measure than those for starting.

Stopper are probably a mixed population, including those near death and those who lose ascertainable medication use through institutionalization.

Their exclusion may represent depletion of susceptibles.
Value of new user designs*

Many studies of drug effects have included prevalent users

This increases sample size, but precludes evaluation of time on study

Effects of a drug may vary by duration of use, e.g. for hormone therapy or coxibs

RCTs control precisely for time on therapy

More value in studying new users

Confounders affecting initiation may differ from those affecting persistence

With prevalent users, need to consider propensity to start and propensity to persist: these may differ

Stopers of preventive drugs are often sick

Risk factors may be affected by use of study drugs, e.g. NSAIDS influence BP, and HT influences CRP

May not be possible to control for these factors on the causal pathway if measured long after initiation
Limitations of new user designs

Reduced power and generalizability through exclusion of long-term users

But long-term users are selected survivors with missing indications for drug initiation

Real-time assessment of determinants of drug initiation may be expensive

But available in computerized databases of medical encounters
Example: non-user vs active referents in a study of NSAIDS & CVD*

Prospective cohort study of CV risks of new use of NSAIDS

Beneficiaries age 65+ in Pennsylvania’s PACE program and in Medicare

Exposed subjects had no NSAID Rx for 6 months before a new Rx

Referent groups: non-users vs new users of glaucoma drugs or thyroid hormones

*DH Solomon et al. Arth Rheum 2006; 54:1378-89
# Adjusted rate ratios of CVD*

<table>
<thead>
<tr>
<th>Drug</th>
<th>New user referent</th>
<th>Never user referent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RR</td>
<td>CI</td>
</tr>
<tr>
<td>Diclofenac</td>
<td>1.10</td>
<td>0.89-1.37</td>
</tr>
<tr>
<td>Ibuprofen</td>
<td>0.96</td>
<td>0.81-1.14</td>
</tr>
<tr>
<td>Naproxen</td>
<td>0.75</td>
<td>0.62-0.92</td>
</tr>
</tbody>
</table>

*From a proportional hazards model; outcome is new MI or stroke*
Design approaches for healthy users

0) Incident and prevalent drug users vs. non-users (matched by exact date)

1a) Incident drug users vs. non-users (matched by exact date)

1b) Incident drug users vs. non-users (matched by date and system use)

2) Incident drug users vs. incident comparison drug users

3) Incident drug users vs. incident comparison drug users without contraindications

4) Adherent incident drug users v. adherent incident comparison drug users without contraindications

Restrict to incident drug users

Match non-users on system use

Restrict to incident comparison drug users

Restrict to patients w/o contraindications

Restrict to adherent patients

Restrict to RCT inclusion criteria

RCT population

S. Schneeweiss et al., Med Care 2007
Impact of restriction on hazard ratios

A) Unadjusted mortality rate ratio estimates:

B) Multivariate adjusted mortality rate ratio estimates:
Limitations of never user referents

Non-users may have different utilization of prescription programs and physicians

Surveillance of their health may be less complete

Covariates may be measured with different errors than regular users of programs

Such confounding may not be controllable
In a new user study, choice of an appropriate active referent group can be challenging.

The comparison treatment will have its own efficacy and safety profile which can complicate interpretation.

However, individuals initiating that alternative will be active within the system, willing to start a new therapy, and screened for its appropriateness.

In many circumstances, inferences from such a comparison will be more valid than those obtained from a passive, non-user referent group.
Summary confounder scores in study design

Randomized trials use eligibility (exclusion) criteria to target patients most likely to benefit and limit side effects. Often leads to exclusion of older, sicker patients, leaving clinicians uncertain about effectiveness in these patients. For example, seminal trial of oxaliplatin for colon cancer (T Andre et al; NEJM 2004) excluded patients >75 years old, yet the disease occurs frequently in this group.

CER can use disease risk scores and propensity scores to target populations at reasonable risk and with treatment uncertainty in whom valid comparisons are possible.
Value of summary confounder scores*

Reduce the frequently high dimension of confounding variables

Both propensity scores and disease risk scores have inherent scientific interest in studies of drug use and outcomes

Both scores provide a natural scale for evaluation of effect modification

Both provide a ready way to identify areas of non-overlap between treatment groups

*OS Miettinen. Am J Epidemiol 1976
Value of summary confounder scores*

While propensity scores (and perhaps also disease risk scores) are often seen as analysis tools, they can (and should) play a role in study design.

Propensity score matching (as well as stratification and weighting approaches) focus on similar subjects who might receive either treatment.

Propensity score development and subject selection can occur before consideration of outcome data.

*DB Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med 2007
Examples of disease risk scores

Framingham risk score for prediction of a first major coronary event

Gail model for prediction of incident breast cancer

APACHE II score for ICU survival

Charlson comorbidity index predicting mortality from administrative data
Challenges to propensity score estimation

Sometimes we do not have large numbers of exposed and unexposed subjects for propensity score estimation.

E.g. when studying newly marketed drugs; or unusual (e.g. high) doses; or multi-category exposures.

Propensity scores may evolve with time, especially after introduction of a new therapy.

Strong predictors of exposure without effect on the outcome should be excluded (MA Brookhart et al; AJ E 2006), and these are difficult to identify in the above situations.
Challenges to estimation of disease risk scores

The disease risk score is best estimated in the unexposed: Problematic with common exposures and when several active referents are compared

Balance is only evaluable in the unexposed

Estimation in the study population can lead to bias; historical controls may provide the best source for estimation of disease risk (prognostic) scores

BB Hansen. Biometrika 2008
Joint matching on the propensity score (PS) and disease risk score (DRS)

Straightforward to minimize the distance between (PS, DRS) pair in exposed and unexposed subjects

With two (or more) exposed groups (versus standard treatment), can minimize distance in three dimensions (PS1, PS2, DRS)

With common exposures and less data on outcomes, emphasize (overweight) distance on the PS dimension

With new or rare treatments, emphasize DRS dimension
Settings favoring disease risk scores

Initiatives such as the FDA’s sentinel system seek to identify adverse effects and benefits of drugs and new doses as soon after marketing as possible.

In the short-term after marketing, propensity scores may be unreliable or evolving.

Historical data on outcomes in people treated before use of the new drug or dose can characterize risk.
Example: new statins & doses after heart attack*

The randomized 4S (1994), CARE (1996) and LIPID (1998) trials proved the value of statin therapy after myocardial infarction (MI) for prevention of recurrent MI, stroke and death

Later trials (PROVE-IT TIMI 22 [2004]; TNT [2005]) showed that higher statin doses (with atorvastatin) yielded greater benefits

We considered disease risk scores to account for confounding in observational data on atorvastatin & statin dose post-MI from 1995-2004

*RJ Glynn et al. Pharmacoepidemiology and Drug Safety 2012
We studied the 5,668 enrollees, aged 65-100 years, in either New Jersey or Pennsylvania’s state-sponsored pharmacy assistance program (PAAD or PACE) who survived an MI and filled a statin prescription within 30 days after discharge from 1995-2004

Exposures of interest are atorvastatin (available starting in 1997) and high-dose statins* (almost never used until 1997)

Outcome was recurrent MI, stroke or death within 1 yr

*NK Choudhry et al. Heart 2007
Prevalence of atorvastatin and high dose statin use
By 6-month time periods from 1995–2004

20 6-month time periods over 10 years
Estimate a disease risk scores

Use data from low-dose, non-atorvastatin users to develop a disease risk score in the period 1995-1997.

Among 826 statin initiators post MI, 203 had recurrent MI, stroke or died within 1 year.

Consider demographic variables, comorbidities, and system use as predictors of outcome.

Logistic model has C-statistic 0.704.
# Predictors of MI/Stroke/Death in 1 yr

<table>
<thead>
<tr>
<th>Variable</th>
<th>Odds ratio</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (1 yr)</td>
<td>1.05</td>
<td>1.02-1.08</td>
</tr>
<tr>
<td>Male</td>
<td>1.2</td>
<td>0.8-1.7</td>
</tr>
<tr>
<td>Black race</td>
<td>2.9</td>
<td>1.5-5.6</td>
</tr>
<tr>
<td>Other race</td>
<td>1.1</td>
<td>0.2-5.7</td>
</tr>
<tr>
<td>NJ vs PA</td>
<td>1.2</td>
<td>0.8-1.7</td>
</tr>
<tr>
<td>MI hosp. 5-6 days</td>
<td>1.2</td>
<td>0.7-2.1</td>
</tr>
<tr>
<td>MI hosp. 7-9 days</td>
<td>1.4</td>
<td>0.8-2.5</td>
</tr>
<tr>
<td>MI hosp. 10+ days</td>
<td>1.5</td>
<td>0.8-2.6</td>
</tr>
</tbody>
</table>
# Predictors of MI/Stroke/Death in 1 yr

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Odds Ratio</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angioplasty</td>
<td>0.6</td>
<td>0.4-0.8</td>
</tr>
<tr>
<td>Cong. heart failure</td>
<td>1.7</td>
<td>1.2-2.5</td>
</tr>
<tr>
<td>Hx stroke</td>
<td>0.8</td>
<td>0.6-1.2</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.8</td>
<td>0.5-1.2</td>
</tr>
<tr>
<td>Hypertension</td>
<td>0.7</td>
<td>0.5-1.0</td>
</tr>
<tr>
<td>Prior MI</td>
<td>0.9</td>
<td>0.5-1.7</td>
</tr>
<tr>
<td>Peripheral vasc. dis.</td>
<td>1.3</td>
<td>0.9-2.0</td>
</tr>
<tr>
<td>Chronic kidney dis.</td>
<td>1.1</td>
<td>0.7-1.7</td>
</tr>
<tr>
<td>Prior hosp. days</td>
<td>1.01</td>
<td>0.99-1.03</td>
</tr>
<tr>
<td>Charlson score (per pt)</td>
<td>1.2</td>
<td>1.1-1.4</td>
</tr>
</tbody>
</table>
Distribution of the disease risk score
Comparison of atorvastatin users and non-users

Number of patients: 3318
- Median: 0.237
- Mean: 0.277
- Standard deviation: 0.168

Number of patients: 1851
- Median: 0.235
- Mean: 0.272
- Standard deviation: 0.166
Distribution of the disease risk score
Comparison of high versus lower-dose statin users

- **Number of patients**: 4247
- **Median**: 0.237
- **Mean**: 0.276
- **Standard deviation**: 0.167

- **Number of patients**: 922
- **Median**: 0.235
- **Mean**: 0.272
- **Standard deviation**: 0.167

Based on the disease risk score

**Predicted Probability: stroke_mi_death = 1**
## Patient characteristics: 1997-2004

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Atorva. (n=1851)</th>
<th>Other statin (n=3318)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, mean sd</td>
<td>78.2 6.6</td>
<td>78.7 6.5</td>
</tr>
<tr>
<td>Male, %</td>
<td>24.3</td>
<td>26.4</td>
</tr>
<tr>
<td>Black race, %</td>
<td>5.6</td>
<td>5.8</td>
</tr>
<tr>
<td>Angioplasty, %</td>
<td>62.9</td>
<td>61.2</td>
</tr>
<tr>
<td>Hyperten., %</td>
<td>80.5</td>
<td>81.0</td>
</tr>
<tr>
<td>CHF, %</td>
<td>59.9</td>
<td>59.9</td>
</tr>
<tr>
<td>Charlson score</td>
<td>2.6 2.0</td>
<td>2.5 2.0</td>
</tr>
<tr>
<td>Disease risk</td>
<td>26.4 15.2</td>
<td>27.0 15.7</td>
</tr>
</tbody>
</table>
## Patient characteristics: 1997-2004

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Hi dose (n=922)</th>
<th>Standard dose (n=4247)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, mean sd</td>
<td>77.7 ± 6.5</td>
<td>78.7 ± 6.5</td>
</tr>
<tr>
<td>Male, %</td>
<td>24.8</td>
<td>25.8</td>
</tr>
<tr>
<td>Black race, %</td>
<td>6.5</td>
<td>5.5</td>
</tr>
<tr>
<td>Angioplasty, %</td>
<td>63.9</td>
<td>61.3</td>
</tr>
<tr>
<td>Hyperten., %</td>
<td>83.1</td>
<td>80.3</td>
</tr>
<tr>
<td>CHF, %</td>
<td>61.4</td>
<td>59.5</td>
</tr>
<tr>
<td>Charlson score</td>
<td>2.7 2.1</td>
<td>2.5 2.0</td>
</tr>
<tr>
<td>Disease risk</td>
<td>26.2 15.3</td>
<td>26.9 15.6</td>
</tr>
</tbody>
</table>
Atorvastatin, dose & outcomes: 1997-2004

After 1996, 5169 post-MI patients initiated statin therapy, 1163 (22.5%) of whom had another MI, stroke, or died within 1 year.

Those without an event had median predicted risk of 0.217 (IQR: 0.137-0.328)

Those with an event had median predicted risk of 0.308 (IQR: 0.197-0.434)
Atorvastatin use and outcomes: 1997-2004

Atorvastatin use was associated with an 8% reduction in the odds of MI/stroke/death within 1 yr (OR 0.92; 95% CI: 0.80-1.05; P =0.23)

After control for the estimated disease risk score, atorvastatin use was associated with a 6% reduction in the odds of MI/stroke/death within 1 yr (OR 0.94; 95% CI: 0.81-1.08; P =0.36)
High dose statin and outcomes: 1997-2004

Use of a high dose statin was associated with a 7% reduction in the odds of MI/stroke/death within 1 yr (OR 0.93; 95% CI: 0.78-1.11; P =0.41)

After control for the estimated disease risk score, high-dose statin use was associated with a 5% reduction in the odds of MI/stroke/death within 1 yr (OR 0.94; 95% CI: 0.80-1.13; P =0.56)
Possible effect modification by DRS quintiles

<table>
<thead>
<tr>
<th>DRS Quintile</th>
<th>Observed risk % (events/exp.) Atorvastatin</th>
<th>Other statin</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.6 (48/381)</td>
<td>10.4 (68/652)</td>
<td>1.21 (0.9-1.7)</td>
</tr>
<tr>
<td>2</td>
<td>15.7 (59/375)</td>
<td>17.8 (117/659)</td>
<td>0.89 (0.7-1.2)</td>
</tr>
<tr>
<td>3</td>
<td>19.2 (71/369)</td>
<td>21.8 (145/665)</td>
<td>0.88 (0.7-1.1)</td>
</tr>
<tr>
<td>4</td>
<td>29.5 (108/366)</td>
<td>27.5 (184/668)</td>
<td>1.07 (0.9-1.3)</td>
</tr>
<tr>
<td>5</td>
<td>31.4 (113/360)</td>
<td>37.1 (250/674)</td>
<td>0.85 (0.7-1.0)</td>
</tr>
</tbody>
</table>
Propensity score limitations

Fitting a propensity score in early follow-up was problematic:

During 1997-98, only 68 subjects used high-dose statins

Propensity score fitted in this period with the 20 covariates of interest showed evidence of over-fitting

Coefficients in this PS had large differences from estimates in a model fitted to data from 1999-2004

Unclear whether differences represented sampling variability or true changes in treatment preferences
Treatment effect heterogeneity across PS strata

Apparently different treatment effects in the tails of the PS can substantially influence estimated treatment effects, possibly due to unmeasured confounding.

Example: T Kurth et al (AJE 2006) found raised death rates only in the subgroup of stroke patients who received TPA (vs untreated) with a low PS for treatment.

Also, M Lunt et al (AJE 2010) found raised death rates in rheumatoid arthritis patients who received biologics (vs untreated) with a low PS, but a protective effect among those with a high PS.
Trimming to the population with equipoise

Propensity scores can also aid in the identification of the individuals with reasonable probability of receiving either treatment of interest. Trimming of subjects in both tails of the propensity score can reduce bias under a range of reasonable scenarios. Treatment effect estimates based on such range restrictions do not correspond to a causal parameter, but may be less prone to bias from unmeasured confounding.

T Stürmer et al, Am J Epidemiol 2009
Trimming to the population with equipoise
Design issues in 3-way CE studies

Comparison of safety and effectiveness of 3 alternative treatments involves special considerations

For example, DH Solomon et al (Arch Intern Med 2010) compared 3 analgesics (nonselective NSAIDS, coxibs, and opioids) in arthritis patients in a new-user design.

Solomon et al created two 1-1 matches with an nsNSAID patient matched to an opioid initiator and separately to a coxib initiator, based on 2 PSs with nsNSAIDs as the referent.
Design issues in 3-way CE studies

The design of Solomon et al evaluates patients with comparable propensities to all three analgesics.

An alternative approach uses a single generalized propensity score based on multinomial logistic regression, and forms 1:1:1 matches for comparison (J A Rassen et al. Epidemiology 2013). Simulations suggest that this alternative may reduce bias.

Separate comparisons of each of the 3 pairs of treatments would include more subjects but would not identify the best treatment for a patient who could reasonably get any of the 3 treatments.
Hypothesis formulation in CE studies

Should design of CE studies proceed from the perspective of a superiority or an equivalence trial?

With approved drugs, perhaps the assumption of equivalence is reasonable.

Noninferiority trials often require large sizes.

CE studies can consider multiple endpoints; e.g. Solomon et al (Arch Intern Med 2010) considered 14 individual and 6 composite endpoints in 3 analgesic treatment groups.
Hypothesis formulation in equivalence trials

Exact equivalence (the null hypothesis) cannot be proven.

Specify $\delta$ such that smaller differences are clinically inconsequential.

Here $H_0$: differences are less than $\delta$; vs $H_a$: differences are at least $\delta$.

Under $H_0$, the upper $100(1-\alpha)\%$ CI will not exceed $\delta$ with probability $1-\beta$.

Need $\beta$ small to ensure a new therapy is allowed to replace an older standard (LM Friedman et al. Fundamentals of Clinical Trials; 3rd edition).
Example: relative efficacy in the VALIANT trial

The Valsartan in Acute Myocardial Infarction (VALIANT) trial compared the efficacy of valsartan (V) with that of captopril (C) in patients with MI complicated by LV systolic dysfunction and/or heart failure*

Patients were randomized to valsartan alone, captopril alone, or both with a 1:1:1 allocation ratio

Primary aims were superiority comparisons of V alone vs C alone, and V plus C vs C alone; focus on all deaths

In absence of superiority, focus on equivalence

Example: setting $\delta$ in the VALIANT trial

VALIANT randomized 14,703 patients to accrue 2,700 deaths so as to have 86% power to detect a 15% reduction in risk, separately for the V vs C and V+C vs C comparisons*

Noninferiority evaluation done conditional on a null result for superiority of V vs C

Choose $\delta$ from a meta-analysis of post-MI ACE trials (SAVE, TRACE, AIRE); pooled OR for death: 0.74

Upper bound of 1.13 for V vs C HR preserves 55% of ACE survival benefit; trial had 74-88% power to find V as effective as C

Alternative hypothesis formulation

With $n$ treatments compared, each with success (or adverse effect rate) $R_i$, $i=1,\ldots,n$

Let $R_{\text{max}} = \max(R_1,\ldots,R_n)$ and $R_{\text{min}} = \min(R_1,\ldots,R_n)$

Would a more relevant study goal, achievable with fewer subjects, be to identify the most or least effective treatment with high probability?

e.g. $\Pr(R_i=R_{\text{max}})>0.975$ or $\Pr(R_i=R_{\text{min}})>0.975$?