# Defining "baseline" using propensity score matching: Application to a clinical trial

Elizabeth Stuart

Johns Hopkins Bloomberg School of Public Health
Department of Mental Health
Department of Biostatistics

Samantha R. Cook, Google
Donald B. Rubin, Harvard University

January 18, 2008

# Outline

# Outline

# Overview of substantive problem

- FDA clinical trial of enzyme replacement therapy for rare disease (Fabry disease)
  - No previously existing treatment
- Because of excellent short-term results, new treatment became open-label during trial...some control patients may start receiving the new treatment
- But interest also in continuing trial and estimating longer-term effects....how?
  - Standard analysis would simply treat data as censored
  - Instead, use historical patient information to help impute outcomes for controls that started taking the drug, as if they had not done so

# Two issues

1. How to ensure that the historical patients are similar to those in the trial?
2. How to define "baseline" for those historical controls?
   - To measure "covariates" and "outcomes", need to identify the date they looked like they could have enrolled in the trial
   - Related to any longitudinal study where treatment assignment date is undefined for control group (e.g., effects of drug use on depression levels 6 months later, effects of being arrested on later criminal activity)

# Three steps

1. Define the point in time when each historical patient looked the most similar to a patient in the trial–their "baseline"
   - Need to identify a particular point in time–matching easiest way to do that
2. Keep those historical patients whose baseline values look the most similar to the patients in the trial
   - For some historical patients, at their baseline they might not actually look very similar to the trial patients
   - Could use weighting or subclassification at this point, but matching offers some advantages in terms of exposition and clarity
3. Generate reasonable model of outcome of interest (serum creatinine) using those historical patients and impute missing outcomes for the trial control patients who switched
   - Let trial patients provide information on short-term trends
   - Historical patients provide information on long-term trends

# Data summary

Two data sets:

1. Double-blind randomized trial
   - 72 male patients
   - Monthly measurements for about 3 years
2. Historical data set
   - 447 male patients
   - Up to 15 years of data on each patient
   - 79 patients had at least one observation that met the randomized study criteria and had at least one observation following it
   - Treat each observation as a potential baseline
     - Implies 293 possible versions for matching

Many covariates available in historical data and from clinical trial

- Must be measured the same way in both data sets

# Outline

# Propensity scores

- Ideally would have a historical patient who looks exactly the same as a randomized patient at the time of randomization
- But this difficult in practice
- Instead, match/select using the propensity score, as a summary of all of the covariates
- Propensity score = Probability of receiving treatment, conditional on the covariates: $P(T_i = 1|X)$
- In our case, "treatment"=being in randomized trial

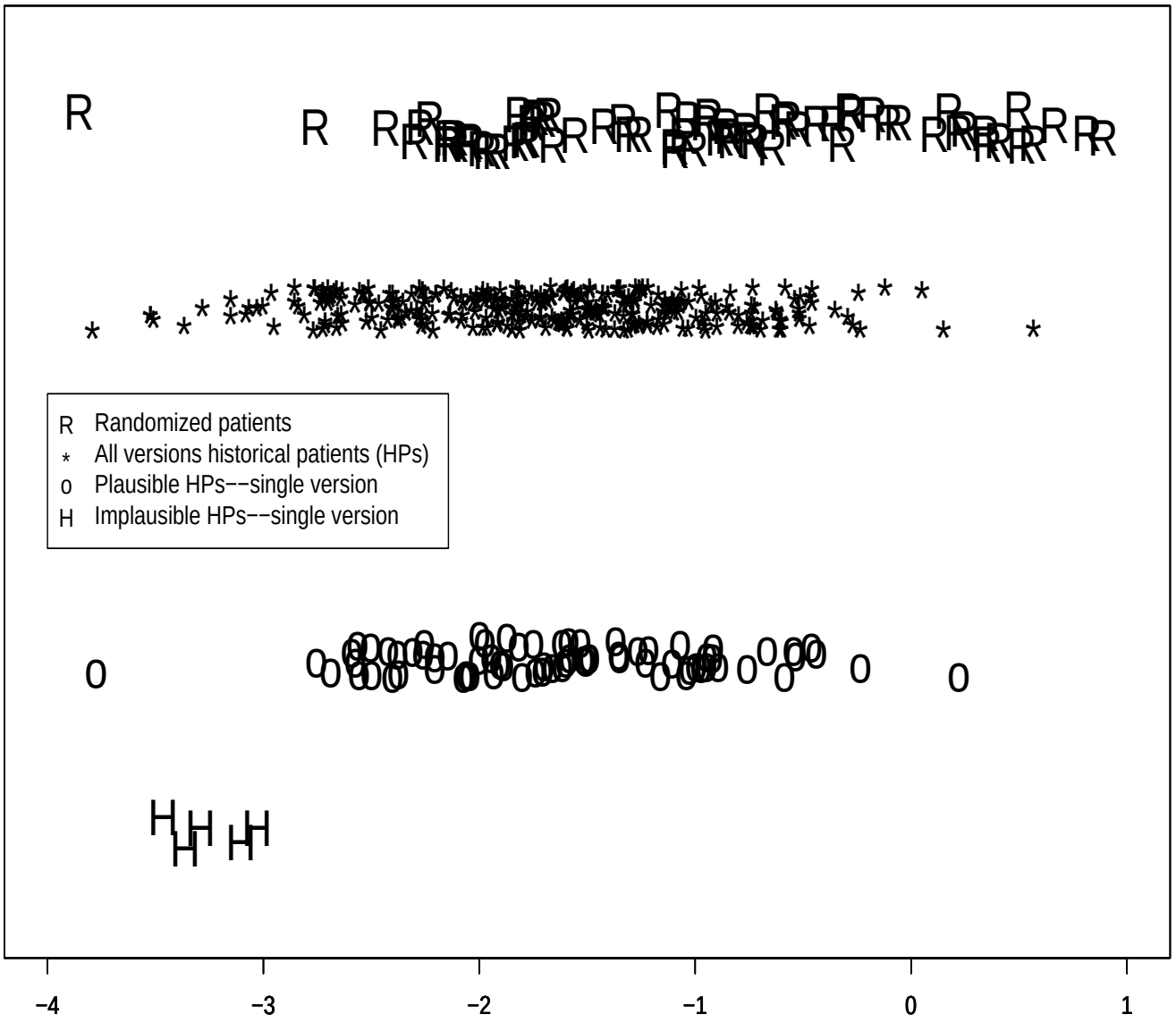
# Propensity score estimation

- Common methods (e.g., logistic regression) require fully observed data
- More complicated if have missing covariate values
- General location model to jointly model categorical and continuous covariates (D'Agostino and Rubin 2000)
  - Treat treatment indicator as one of the categorical variables
  - Have to be careful about which interactions are included
    - As few restrictions as possible on joint distribution of covariates
    - Two-way interactions of all covariates with treatment indicator
  - Fit using ECM algorithm (Schafer 1997)
- Note: lots of open research questions regarding use of propensity scores/matching with missing covariate values

# Selection algorithm

1. Select all "versions" of each historical patient that met enrollment criteria for randomized experiment and that had at least one observation following it
   $\longrightarrow$ 79 patients, 293 possible versions
2. Propensity score estimated on these 293 historical patient versions and 72 randomized patients (Figure 1)
3. For each historical patient, select the version that is closest to a randomized patient. Closest measured by Mahalanobis distance on age and baseline serum creatinine within propensity score calipers.
   - Age and baseline serum creatinine believed to be the most important covariates in terms of predicting the outcome (serum creatinine)
   - Will yield good matches on all of the variables in the propensity score, and particularly good matches on age and serum creatinine

**Figure 1**

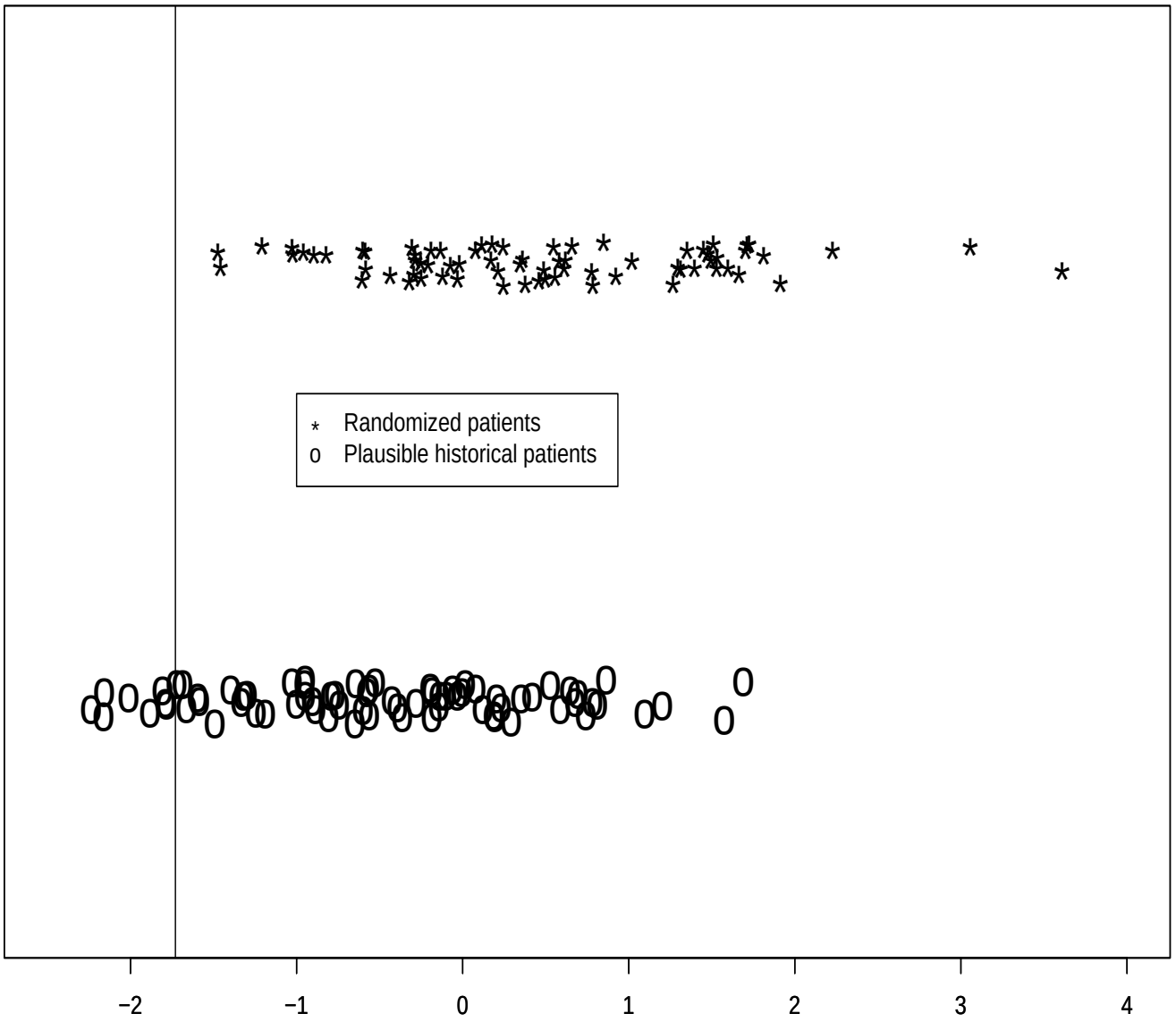

R Randomized patients
* All versions historical patients (HPs)
o Plausible HPs−−single version
H Implausible HPs−−single version
−4
−3
−2
−1
0
1
Logistic Transformation of Propensity Score

4. Discard all other versions of that historical patient
   - If no randomized patient within propensity score caliper, discard that historical patient

   $\longrightarrow$ 74 chosen versions
5. Re-estimate propensity score in randomized versus plausible historical versions (Figure 2)
   - Discard historical patients with propensity scores clearly lower than lowest randomized patient

   $\longrightarrow$ 66 chosen historical versions

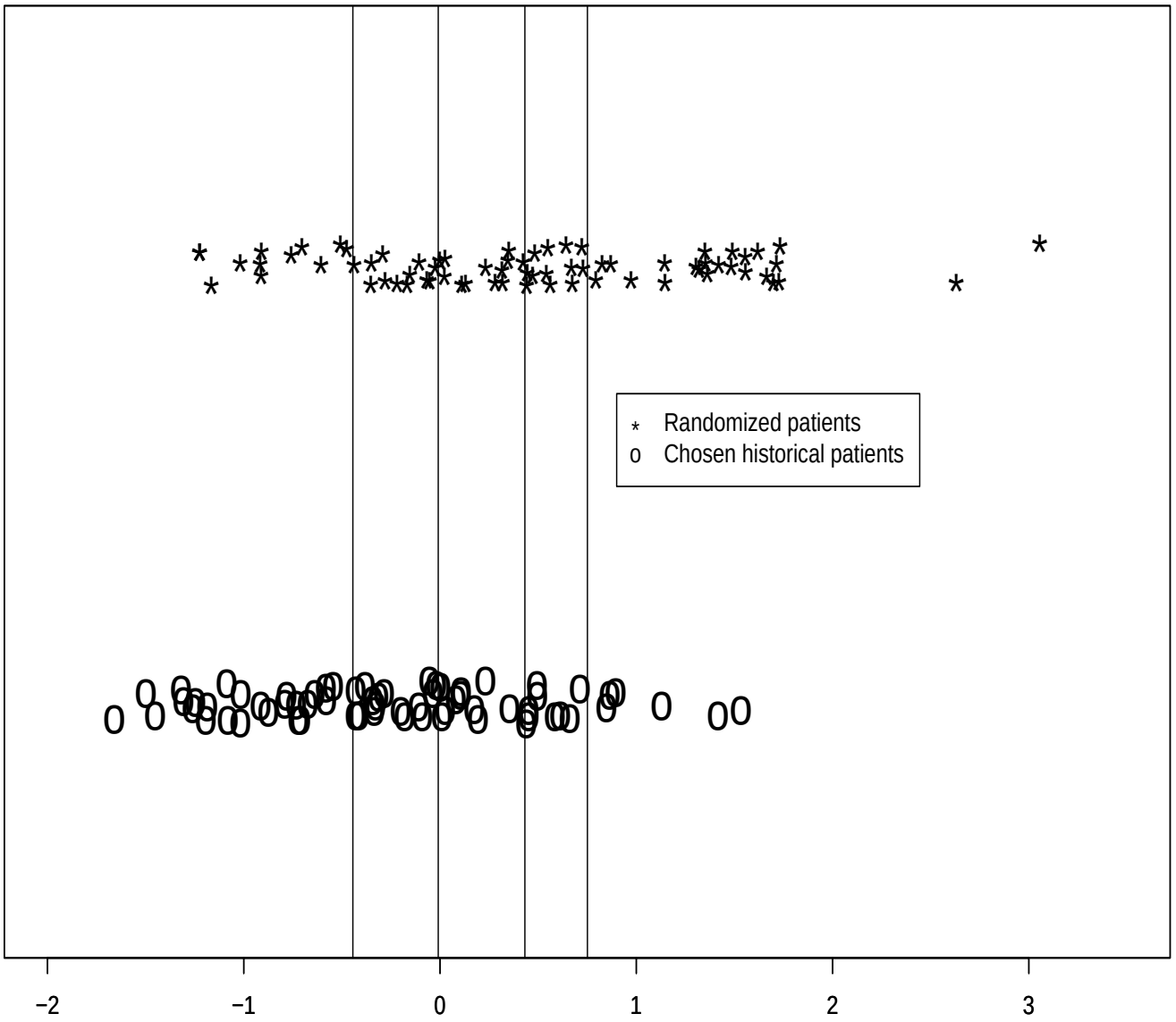
**Figure 2**


* Randomized patients
o Plausible historical patients
−2
−1
0
1
2
3
4
Logistic Transformation of Propensity Score

6. Finally, re-estimate propensity score and subclassify (Figure 3)

- Within subclasses, distribution of covariates approximately the same in the treated and control groups
  - Mini-randomized experiment
- Selected historical patients (and their defined baselines) examined by medical officer to confirm that they looked as if they could have enrolled in the trial
  - One benefit of doing matching (versus weighting or subclassification) is that have this transparent diagnostic
  - Easy for non-statisticians to look at selected historical patients and their baselines and confirm that they look similar to trial patients

**Figure 3**


* Randomized patients
o Chosen historical patients
−2
−1
0
1
2
3
Logistic Transformation of Propensity Score

# Covariate balance in matched samples–fully observed covariates

| | Randomized Mean (SD) | All versions historical | Selected versions historical | Final version (subclasses) |
|---|---|---|---|---|
| N | 72 | 293 | 79 | 66 |
| P score | 0 (1) | 0.69 | 0.71 | 0.19 |
| Age | 45 (9.1) | 0.42 | 0.75 | 0.12 |
| SC | 1.68 (0.51) | -0.14 | 0.33 | -0.01 |
| Estd. GFR | 52.93 (17.8) | 0.07 | -0.41 | -0.01 |
| White | 89% | 0.2% | 5.3% | 2.2% |
| On Ace Inh. | 31.6% | 8.0% | 10.3% | 5.7% |
| Hypertens. | 36.1% | 4.7% | 12.1% | 4.4% |

# Outline

# What happened next?

- End goal: Predictions of outcome (one over serum creatinine) over time for control patients who switched to new therapy, as if they never switched
- Bayesian model of long-term progression in outcome in untreated patients–based on scientific knowledge
  - Quadratic trend in time
  - Constrained to be negative
- Randomized control patients inform short-term (linear) trends
- Historical patients inform long-term (quadratic) trends

# Steps in doing the imputations

1. Model of long-term trends in outcome fit using matched historical patients
2. Obtain posterior distribution of quadratic coefficients from that model
3. Use that posterior distribution as the prior distribution to fit model using randomized control patients
   - First time randomized control group outcomes used
4. Use resulting parameter estimates to predict outcomes for controls who switched to new therapy (using multiple imputation)
5. Estimates of treatment effect estimated using data from trial as well as these imputations
   - First time randomized treatment group outcomes used!

# Outline

# Other approaches?

- What about propensity score weighting or subclassification?
  - Not totally clear how to implement given the two steps of defining baseline and selecting similar individuals
  - Could probably implement at one step at least...e.g., once baseline defined, weight each historical patient by their baseline similarity to trial patients
  - Would not have allowed examination by medical officer quite as easily
  - (In fact, after matches selected, outcome models run with subclass indicators)
- What about getting more than one match?
  - We sort of did this, by allowing all historical patients who had a baseline similar to a trial patient to be included
  - Limited somewhat by sample size: in the end, had about the same number of historical and randomized patients

- Could potentially have instead used "balanced risk set matching" (Li, Propert, Rosenbaum 2001) to select the baseline
  - Would match each patient who received the treatment (was in the trial) at time t to a similar patient who had not yet received the treatment (entered the trial) by time t
  - But not clear if it would work in this setting...requires common time scale of measurements for all patients (?)

# Other issues that came up...topics for further research

- Use and diagnosis of propensity score matching with missing data
  - Guidance for GLOM specification
- Way to include calendar time in matching procedure?
- What were the appropriate cut-offs in terms of historical patients being "close enough"?
  - In our case not much sensitivity, but in other settings there might be

# Conclusions: Implications for trial

- Found group of historical patients who look like they could have been in the randomized trial
- Historical patient information used carefully
  - Only use historical patients similar to those in clinical trial
  - Information from these patients used only as much as necessary (for information on long-term trends)
- Actually, in the end, only 1 control patient switched so methods not necessary!

# Conclusions: Methodology

- Propensity score methods in general offer a few advantages
  - Model hypothetical randomized experiment: ensure using similar individuals
  - Provided way to define baseline for historical patients
  - Well balanced with randomized group
  - No use of outcome (especially important for FDA submission)
  - Allowed us to prioritize good matches on age and serum creatinine
- In our situation, 1:1 matching made the most sense
- Lots of open research questions regarding the use of propensity score methods in practice

# Contact information

- Website: http://www.biostat.jhsph.edu/∼estuart
- Email: estuart@jhsph.edu