`hdps`: Implementation of high-dimensional propensity score approaches in Stata

John Tazare Elizabeth Williamson Ian Douglas

Stata UGM 2019

5th September 2019

@LSHTMstatmethod

Acknowledgements

- This work is funded by the Medical Research Council as part of a Doctoral Training Partnership based at LSHTM.

Outline

Table of Contents

Introduction

- Electronic Health Records (EHRs) increasingly used to investigate the effect of medications
  - Risks/benefits may be different in routine care versus trials
  - EHRs often the best available data to answer these questions
- Invalid results undermine their use
- A key issue is adequate confounder adjustment

Table of Contents

Propensity Scores (PS) in Pharmacoepidemiology

- Models the treatment allocation process
- Defined as conditional probability of being treated given a set of observed covariates
- Typically estimated using logistic regression model
- Methods for estimating treatment effects using PSs include:
  - Covariate adjustment
  - Stratification
  - Matching
  - Inverse Probability of Treatment Weighting (IPTW)

High-Dimensional Propensity Score (hd-PS)

Motivation:

- Absence/imperfect recording of important confounders in EHR data

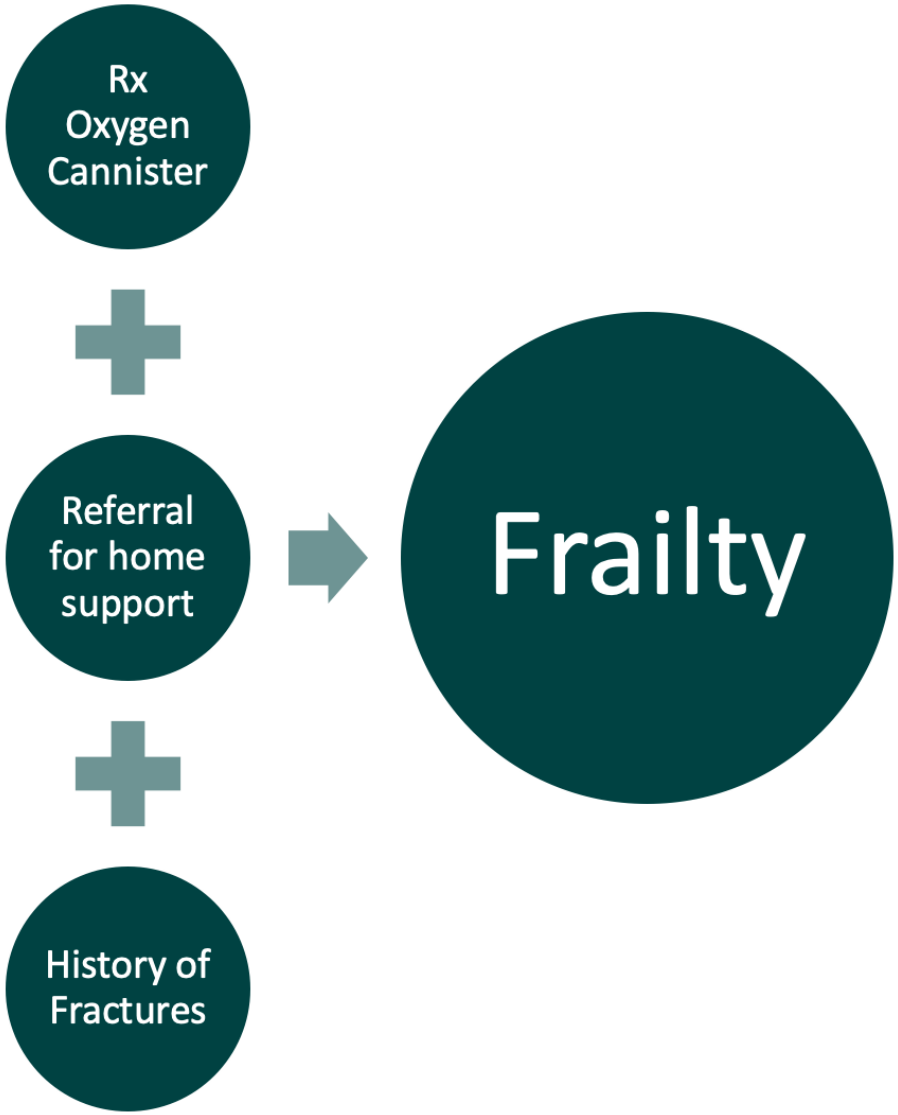
hd-PS:

- Developed in US health claims data [Schneeweiss et al., 2009]
- Information stored as codes in databases are proxies to underlying confounders (or constructs)
- Semi-automated algorithm for selecting confounders

Aim:

- Select important confounders to minimise residual confounding

hd-PS: What do we mean by 'Proxies'?



Description of hd-PS Algorithm

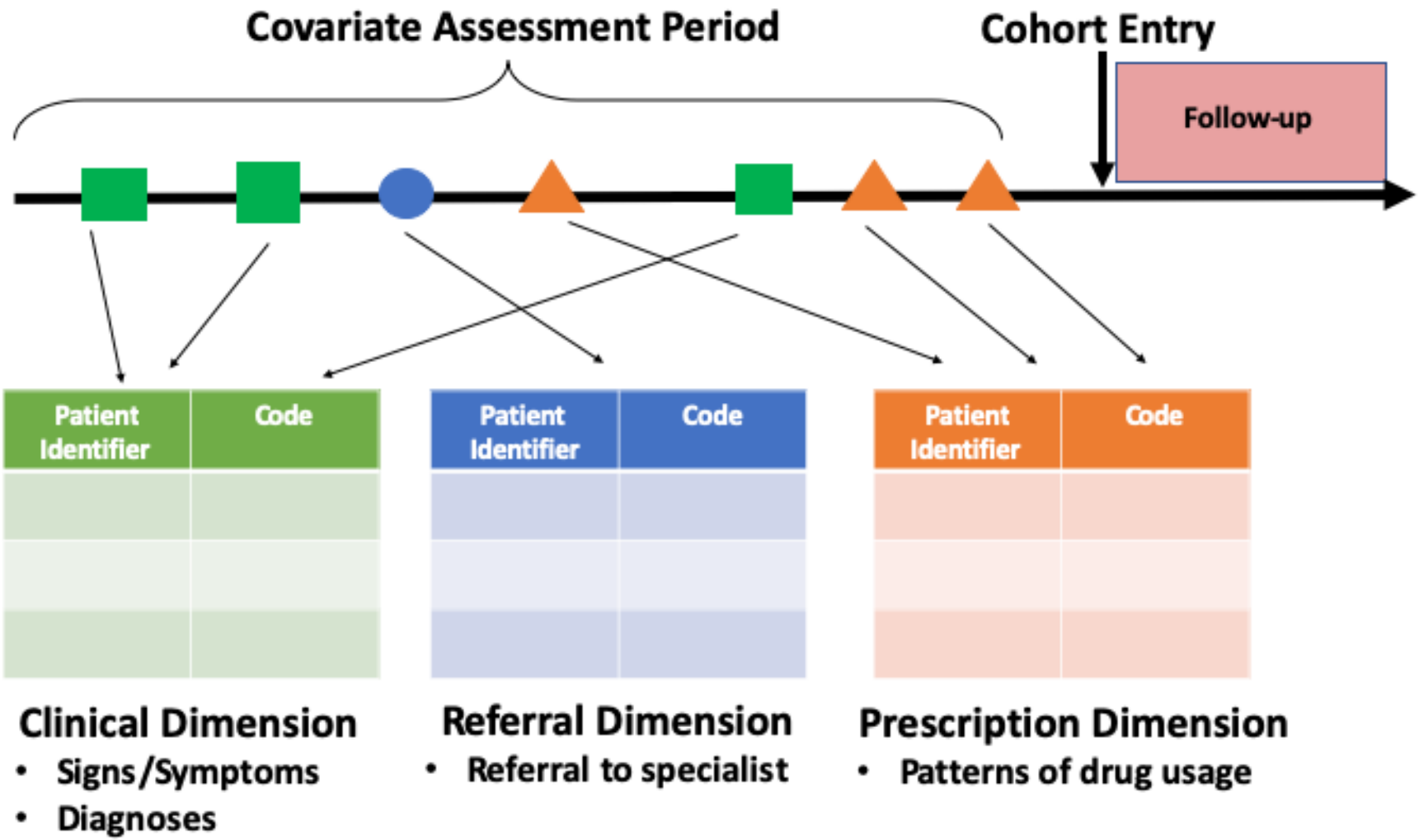
Step 0: Prior to running the algorithm

- Force clinically important factors and demographics into PS model e.g. age, sex and calendar time
- Define a baseline time-window to assess each individual's confounder information

Step 1: Specify a number of data dimensions

- Dimensions represent different aspects of care
- UK EHRs: clinical information, patterns of drug usage and referrals to secondary care

Description of hd-PS Algorithm



Description of hd-PS Algorithm

Step 2: Within each dimension identify the most prevalent codes
(typically $d = 200$)

Step 3: Assess the recurrence of each identified covariate

- 3 indicators of frequency for each code:
  - **Once:** Recorded $\geq$ once for that patient
  - **Sporadic:** Recorded $\geq$ median number of times
  - **Frequent:** Recorded $\geq$ 75th percentile

Description of hd-PS Algorithm

Step 3: Assess the recurrence of each identified covariate

Example: Code=E10 (Type I diabetes)
Median=2
75th percentile=4

Patient	Code Count	E10-Once	E10-Sporadic	E10-Frequent
1	5	1	1	1
2	3	1	1	0
3	1	1	0	0

Description of hd-PS Algorithm

Step 4: Prioritise covariates (within each dimension)

- Covariates with highest potential to bias treatment outcome relationship selected
- Select top empirical candidates from previous step (typically $k = 500$)

Steps 5/6: Perform standard PS analysis

- Estimate treatment PS using predefined and empirically selected variables
- Incorporate PS using standard methods to estimate treatment effect

Table of Contents

hd-PS Software

- hd-PS has been implemented in SAS & R:
  - SAS: www.drugepi.org/dope-downloads/
  - R: github.com/lendle/hdps
- Forthcoming Stata suite: `hdps`
  - Implements traditional hd-PS
  - Extends to hd-PS developments in UK EHRs

`hdps` Suite Overview

- `hdps set`
  - Reads in dimension files
- `hdps prevalence`
  - Must be ran after `hdps set`
  - Step 2: Calculates code prevalences
  - Returns code summary information for codes selected ($d\times$ no. of dims)
- `hdps recurrence`
  - Requires a study cohort dataset in memory
  - Step 3: Recurrence of codes identified by `hdps prevalence` assessed
  - Returns dataset with set of candidate covariates (at most $3 \times d\times$ no. of dims)
  - Step 4: Prioritises covariates and returns dataset with top k

Table of Contents

Case study: Background

Example of contradictory results [Douglas et al., 2012]

- **Population:** Clopidogrel and aspirin users in UK Clinical Practice Research Datalink
- **Treatment:** PPI use vs No PPI use
- **Outcomes:** Myocardial Infarction (MI) analysed using Cox model
- **Findings:**
  - Pattern of associations strongly suggested residual confounding between patients
  - Self-controlled case series - no evidence of increased risk
  - Subsequent trials/genetic studies confirmed lack of association

Case study: Methods

Re-analysis of original study:

- PS analysis adjusting for the original confounders
- Confounders:
  - Age, sex, smoking status, alcohol consumption, BMI categorised, diabetes, coronoary heart disease, peripheral vascular disease, ischaemic stroke, and cancer
- PS incorporated using inverse probability of treatment weighting (IPTW)

Case study: Methods

hd-PS analysis:

- Identified 3 dimensions: Clinical, Referral, Prescription
- 200 most prevalent variables chosen from each dimension
- 500 variables added to PS model + original confounders

Aim:

- Obtain a point estimate closer to the expected null result with similar precision to the original study

Case study: Results

Analysis	HR (95% CI)
Original Analysis	
Crude	1.23 (1.06 – 1.42)
Investigator	1.17 (1.00 – 1.35)
hd-PS Analysis	
hd-PS Adapted for UK EHR	1.00 (0.78 – 1.28)

Case study: Results

Analysis	HR (95% CI)
Original Analysis	
Crude	1.23 (1.06 – 1.42)
Investigator	1.17 (1.00 – 1.35)
hd-PS Analysis	
hd-PS Adapted for UK EHR	1.00 (0.78 – 1.28)

Conclusion

- hd-PS improved adjustment for confounding compared with traditional methods
- Captured extra predictors of prescribing which were also causing confounding bias
- Potential to improve confounder adjustment in UK EHRs

Final Thoughts

How best to read/store the dimension files? (datasets vs. matrices)

Thank you for listening

John Tazare
john.tazare1@lshtm.ac.uk
@JohnTStats

References I


Bross, I. (1966).
Spurious effects from an extraneous variable.
J Chronic Dis, 19:637–47.


Douglas, I. et al. (2012).
Clopidogrel and interaction with proton pump inhibitors: comparison between cohort and within person study designs.
BMJ, page e4388.


Schneeweiss, S. et al. (2009).
High-dimensional propensity score adjustment in studies of treatment effects using health care claims data.
Epidemiology, pages 512–22.

A1: Prioritisation using the Bross formula

Step 4: Prioritise covariates (within each dimension)

Defined for binary confounders

$$\text{ARR} = \text{RR} \times \text{bias}_M$$

- ARR: Observed *RR* treatment on outcome adjusted for individual binary confounder (confounded)
- RR: 'Unconfounded' *RR* treatment on outcome

A1: Prioritisation using the Bross formula

Step 4: Prioritise covariates (within each dimension)

$$\text{where} \quad \text{bias}_M = \frac{\text{P}_{C1}(\text{RR}_{CD} - 1) + 1}{\text{P}_{C0}(\text{RR}_{CD} - 1) + 1}$$

- Bross formula [Bross, 1966]
- Strength of confounder on outcome - choose covariates with highest magnitude of bias
- P_{Ci}: Prevalence of binary confounding factor in treated group ($i = 1$) and untreated/comparator group ($i = 0$)
- RR_{CD}: Effect of confounder on outcome