Assessing Treatment Effect Using Propensity Score Matching within the U.K. Population of Crohn's Disease Patients

Laura H. Gunn, Ph.D.
Associate Professor, Public Health Sciences
Director of Health Analytics, College of Health & Human Services
Director, Health Analytics & Outcomes Research Academy
University of North Carolina at Charlotte (UNCC)
&
Honorary Research Fellow
School of Public Health, Faculty of Medicine
Imperial College London

Outline

- **Medical/Population Health Motivation**
- Study Design Issues
- Purpose of Propensity Score Matching (PSM)
- Implementation of PSM & Balance Diagnostics
- Application to Treatment in Crohn's Disease Using CPRD Data
- Next Steps/Other Areas of Application

Population Health Application: Crohn's disease

- > 70% of Crohn's disease (CD) patients have complications within 10 years of diagnosis (Cosnes et al. 2002)
  - ≥ 50% require surgical resection within this time
  - 70-80% require it within lifetime (Loftus 2006)

- Medical treatment needed to reduce surgeries

- Thiopurines (TP)
  - Used in maintenance of remission of CD (Prefontaine et al. 2009)

- Increases in TP use concurrent with falls in surgical resections (Ramadas et al. 2010)

Research Questions/ Medical Practice Evaluation

- ***Evaluate temporal trends in TP prescribing & 1st intestinal resection***
- ***Compare 1st intestinal resection rates in patients treated with and without TP***
- ***When should therapy be initiated?***
- ***For how long should therapy be administered to achieve optimal results in long-term reduction in surgery risk?***

Causal Effect of Thiopurine Treatment on 1st Intestinal Resection in CD

- **Non-randomized study** (longitudinal cohort study)
- U.K. Clinical Practice Research Datalink (CPRD)
  - Over 13 million registered patients with primary care physicians
  - Clinical & prescribing data
  - Practices are regularly audited to ensure data accuracy & completeness
  - Validation studies report high level of inflammatory bowel disease (IBD) recording against medical records (Lewis et al. 2002; 2004)
- CD incident cases diagnosed through 2005 (n=6,159)
  - Patients followed from diagnosis up to 5 years (1989-2010) (registered for ≥ 1 year)
- Excluded patients with:
  - Co-morbid conditions (n=165)
  - Diagnosis with CD at 1st surgery (n=354)
- 5,640 resulting patients

Outline

- Medical/Population Health Motivation
- **Study Design Issues**
- Purpose of Propensity Score Matching (PSM)
- Implementation of PSM & Balance Diagnostics
- Application to Treatment in Crohn's Disease Using CPRD Data
- Next Steps/Other Areas of Application

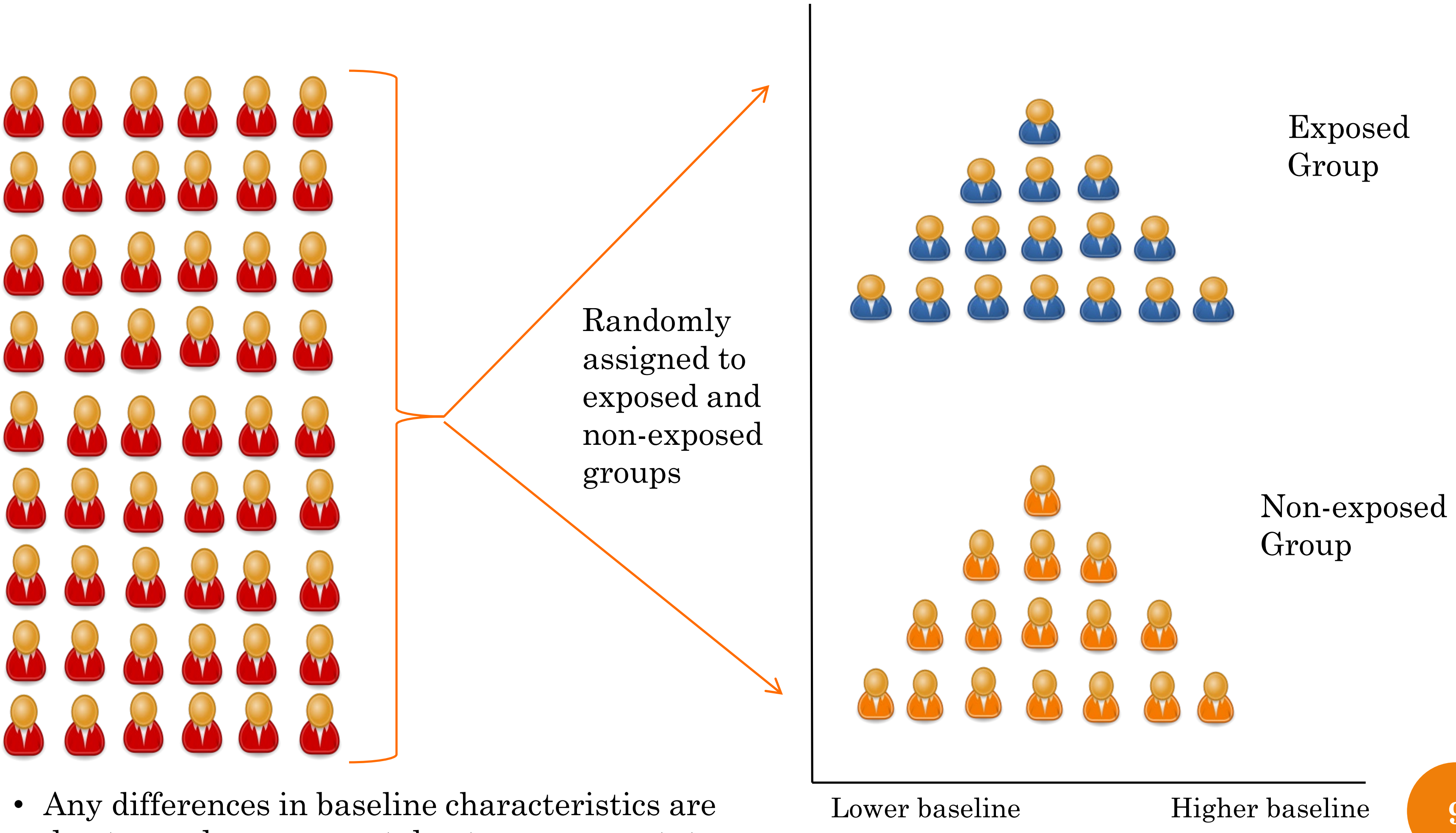
Randomised Controlled Trials (RCTs)

- Gold standard for estimating treatment effects on health outcomes
- Direct comparison of outcomes between intervention & control groups to estimate treatment effect
  - Random allocation of subjects prevents confounding between intervention status & measured baseline characteristics
  - On average, distribution of baseline covariates is similar between intervention & control groups
  - Yields **unbiased** estimate of average treatment effect
    - Continuous data → (Standardized) Difference in Means
    - Binary data → Odds Ratios, Relative Risks, Difference in Proportions
    - Time to Event data → Hazard Ratios

Non-Randomized Studies

- Aim: Estimate a causal effect within an observational study (e.g., not feasible to conduct RCT)
- ***Patient characteristics influence treatment selection***
- Leads to systematic differences in baseline characteristics between exposure & control subjects
  - Produces **biased** estimates of treatment effects
- **Propensity score matching *reduces confounding effects* in observational studies**
  - Creates a pseudo-RCT framework for analysis of exposure effects on outcomes

RCT: Similar Baseline (Distribution) Characteristics by Construction



Randomly assigned to exposed and non-exposed groups

Exposed Group

Non-exposed Group

Lower baseline

Higher baseline

- Any differences in baseline characteristics are due to randomness - not due to exposure status.
- Individuals & their baseline characteristics do not influence exposure status.

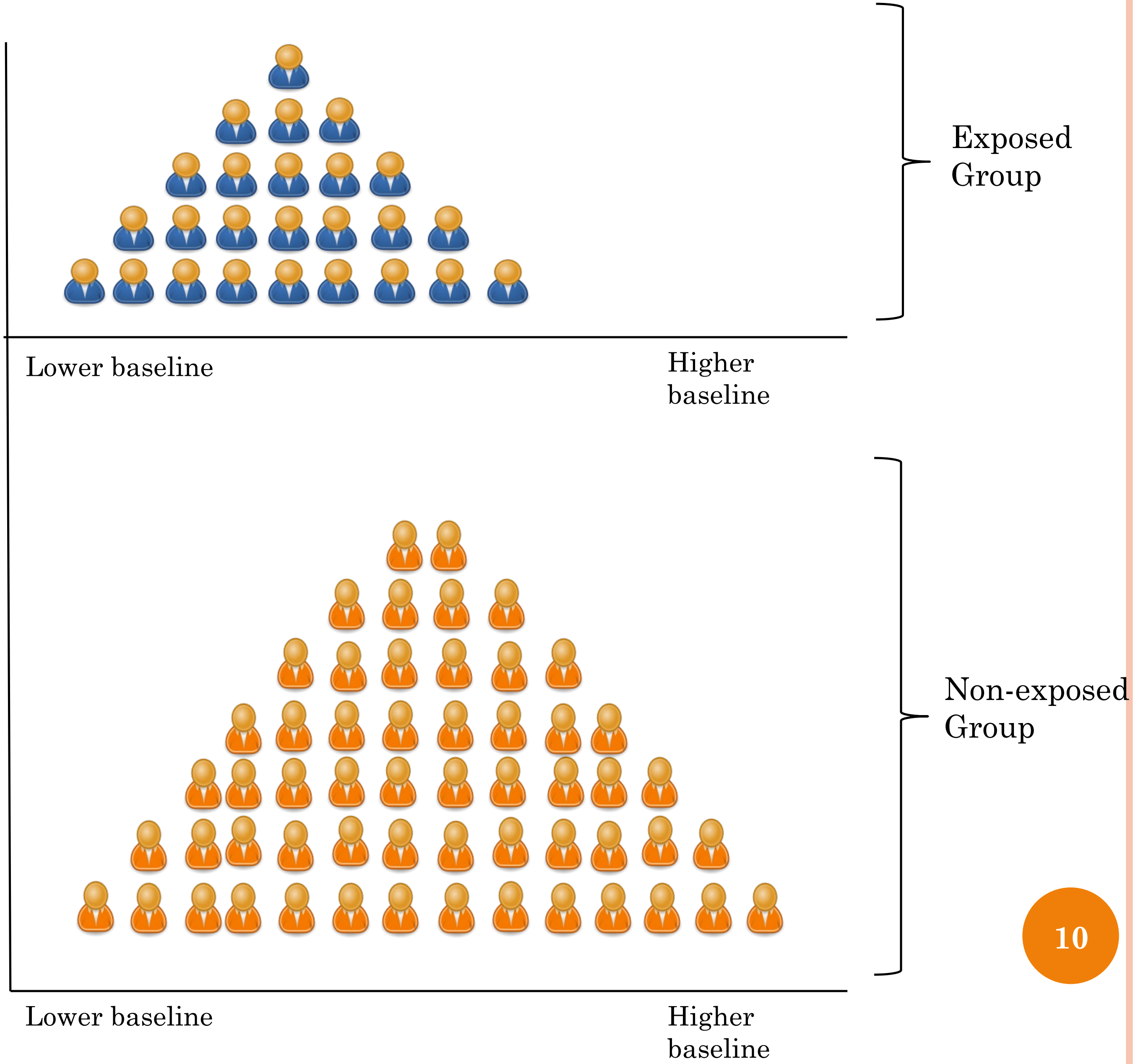
Non-Randomized Studies

- Distribution of baseline characteristics may be different for exposed & non-exposed

- Difference may not occur at random, but determined by exposure/treatment status

- **Crohn's disease patients were not randomized before being exposed/not exposed to TP**

- ***Therefore, exposure status to TP (treatment status) could be affected by baseline characteristics.***



Non-Randomized Studies

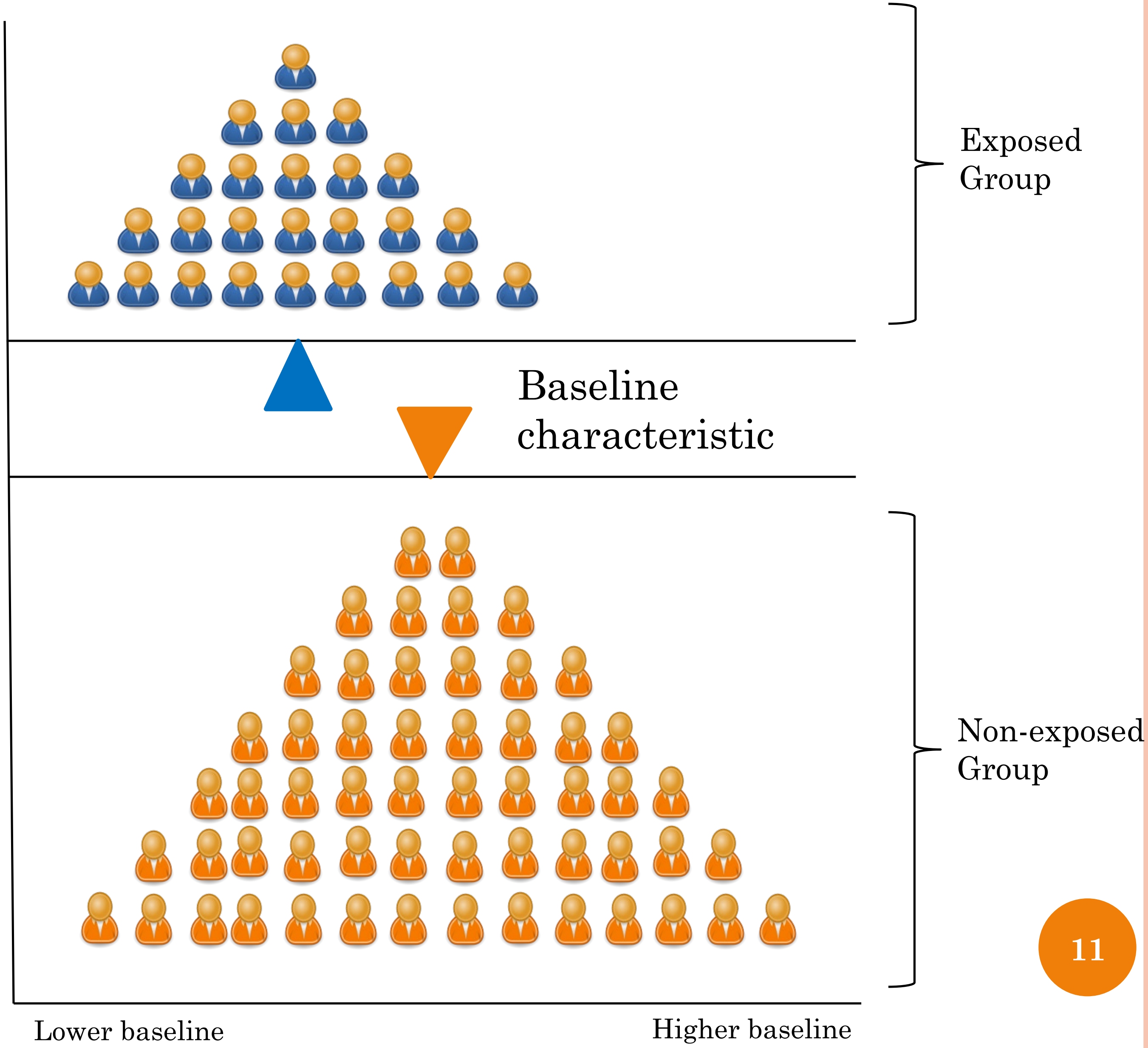
- Different mean levels of a baseline characteristic for the exposed & unexposed groups

- **Baseline characteristic differences can impact differences in average treatment effect**

- For example, the impact of a treatment may be different by age of the treated:

Cannot compare effect of treatment on younger group versus non-treatment on older group.

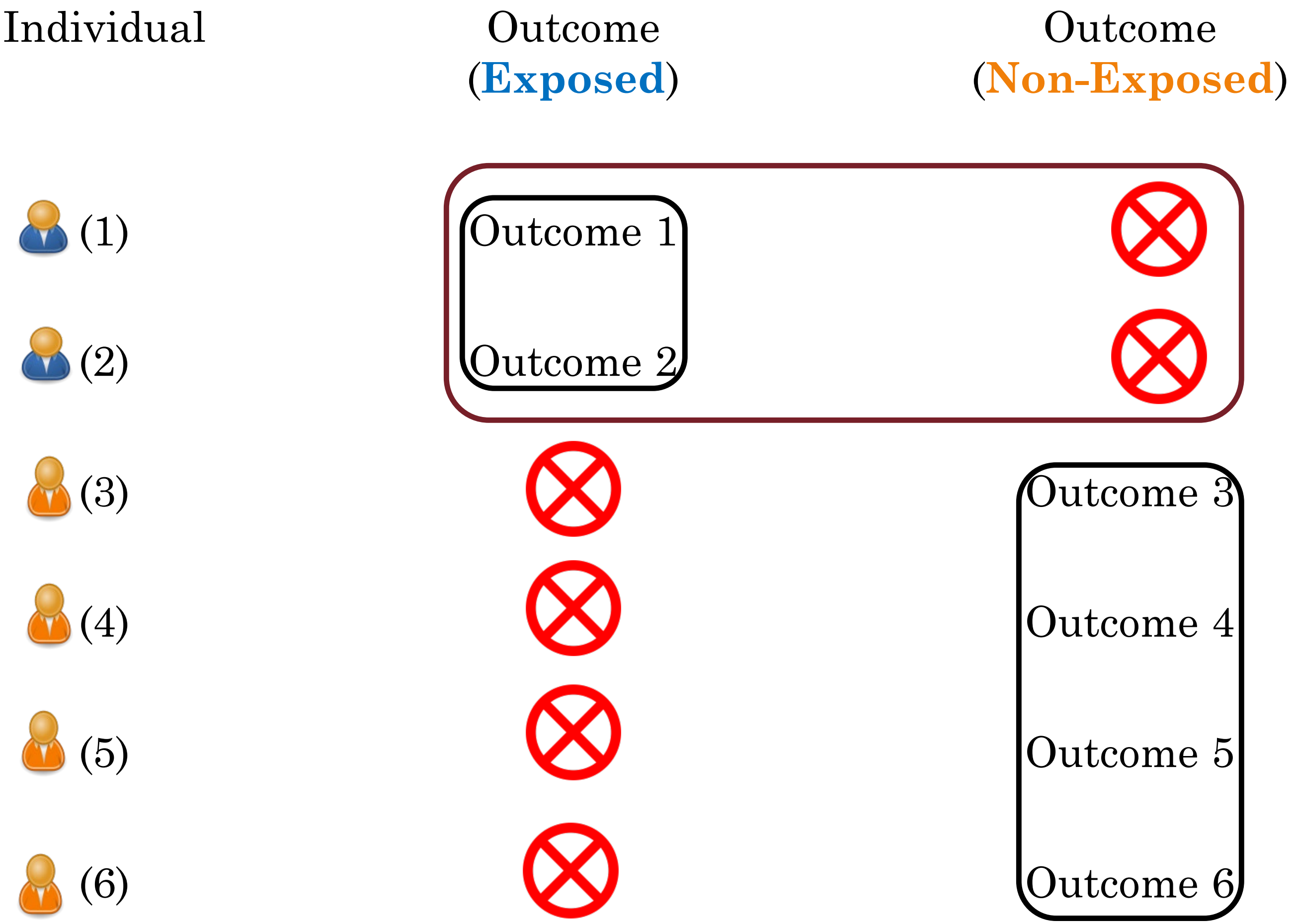
Differences can be due both to age and treatment.



Non-Randomized Studies

- Observational data: what can we estimate?
  - ✓ Average outcome for exposed group
  - ✓ Average outcome for non-exposed group
  - ✗ Differences in outcome *only* due to exposure (i.e., average treatment effect)
- Need non-exposed group to be *similar* to exposed group to assess treatment effect
  - **Propensity score matching methods**
    - **Find a subset of non-exposed individuals who are *similar* to exposed subjects**
    - Estimate the effect of non-exposure only on those individuals
    - Treatment effect: Difference in outcome for exposed vs. matching subset of non-exposed

Non-Randomized Studies

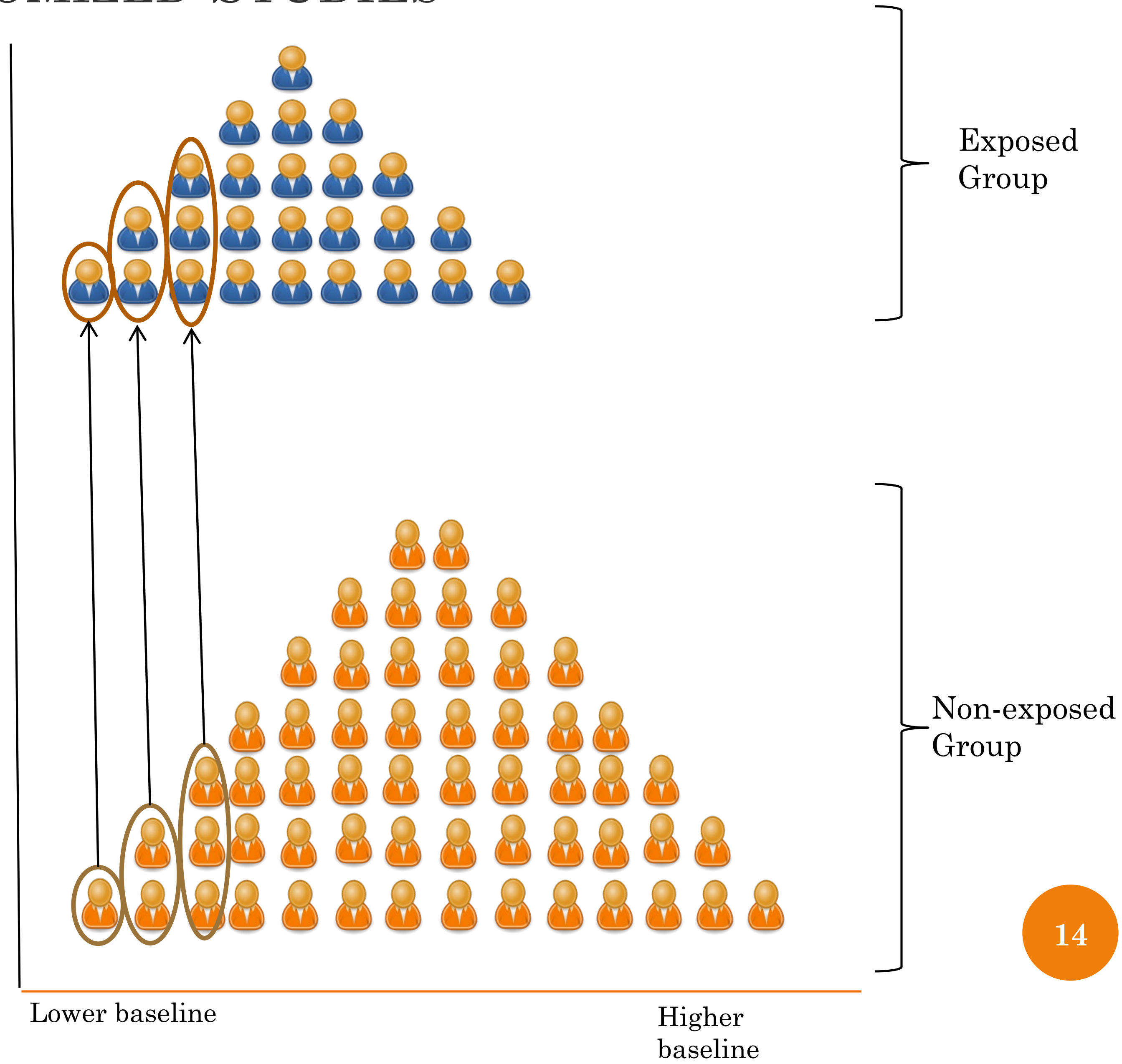


Individual	Outcome (**Exposed**)	Outcome (**Non-Exposed**)
(1)	Outcome 1	⊗
(2)	Outcome 2	⊗
(3)	⊗	Outcome 3
(4)	⊗	Outcome 4
(5)	⊗	Outcome 5
(6)	⊗	Outcome 6

Non-Randomized Studies

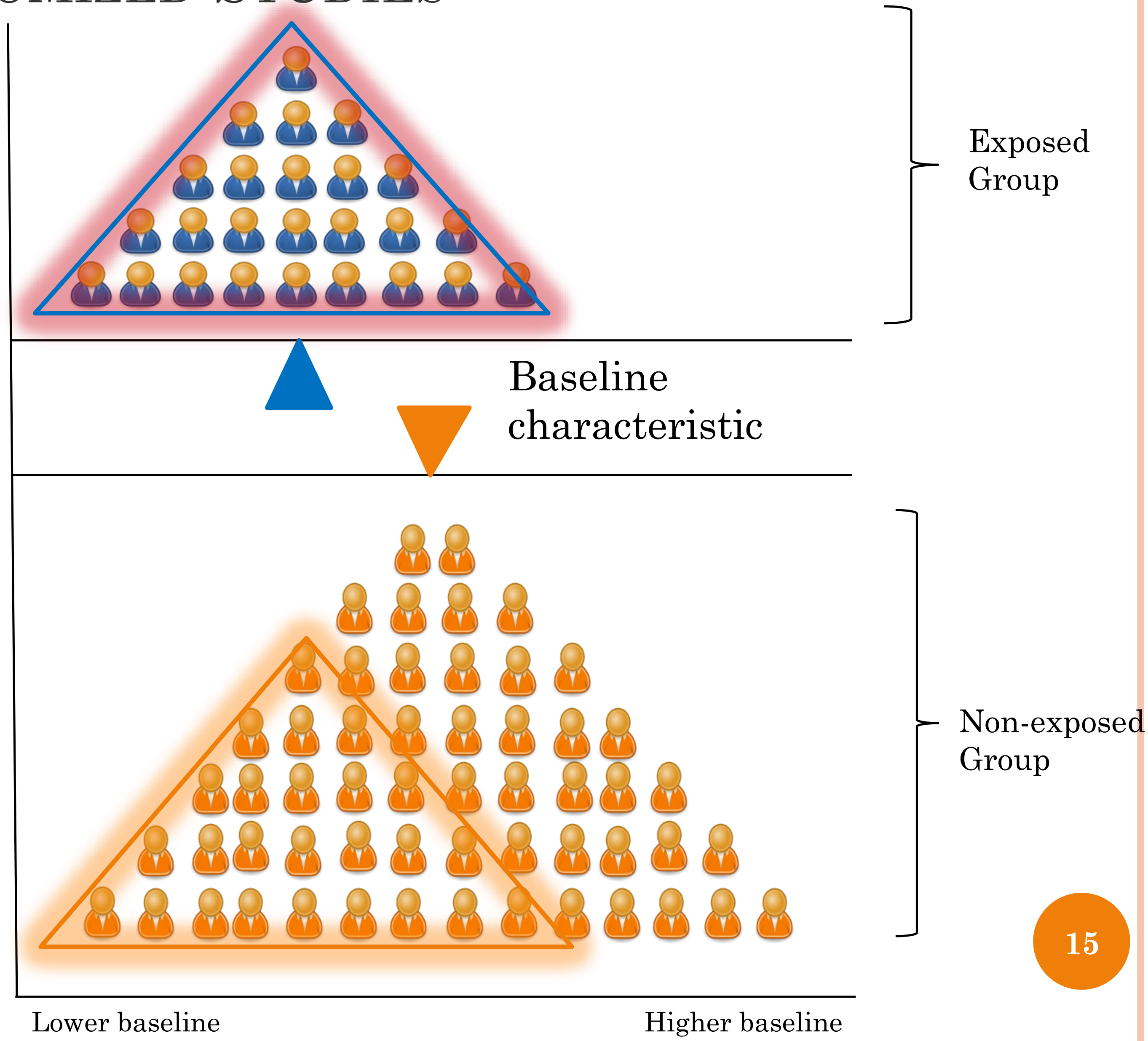
Step 1: Match (i.e., identify) individuals with a similar baseline characteristic.

We find the untreated individuals who best resemble the treated ones by level of the baseline characteristic.



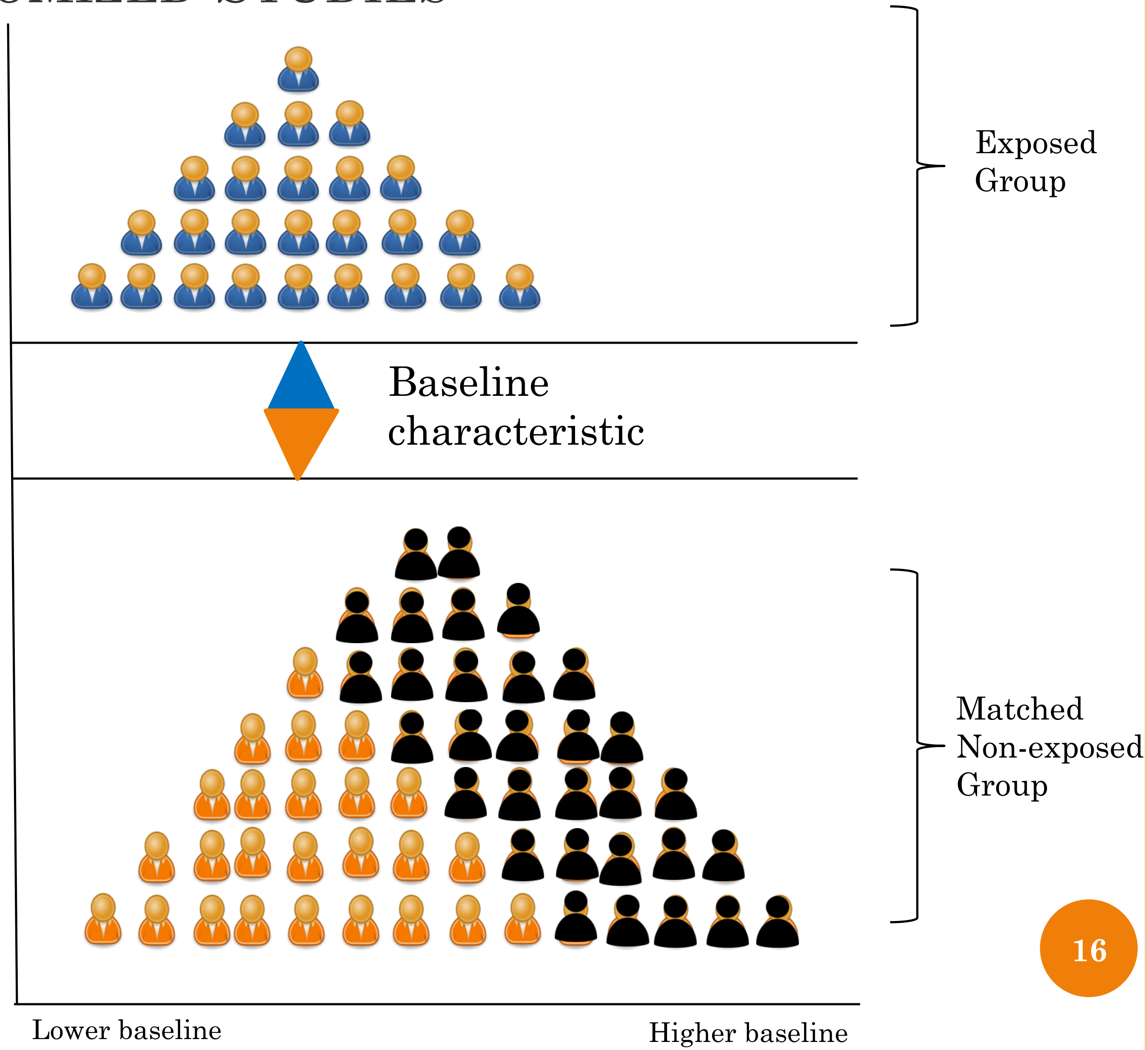
Non-Randomized Studies

Individuals selected from the unexposed group will be the best available matches (by baseline characteristic) for those in the exposed group.



Non-Randomized Studies

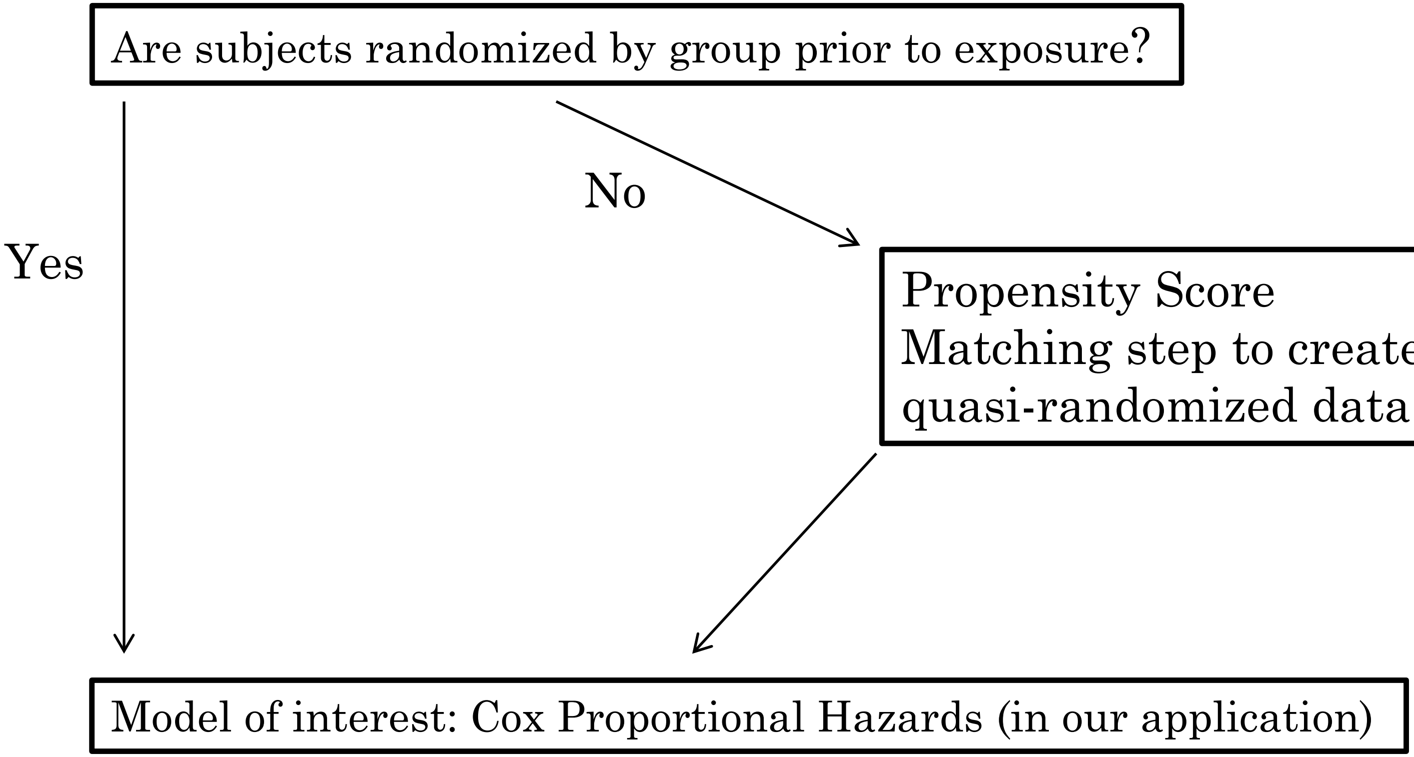
Step 2:
Remove individuals who are not a good match, so that the baseline characteristic has no impact on exposure/treatment status.



OUTLINE

- Medical/Population Health Motivation
- Study Design Issues
- **Purpose of Propensity Score Matching (PSM)**
- Implementation of PSM & Balance Diagnostics
- Application to Treatment in Crohn's Disease Using CPRD Data
- Next Steps/Other Areas of Application

Roadmap to Data Analysis: Causal Effects from Observational Data



Are subjects randomized by group prior to exposure?

Yes

No

Propensity Score Matching step to create quasi-randomized data

Model of interest: Cox Proportional Hazards (in our application)

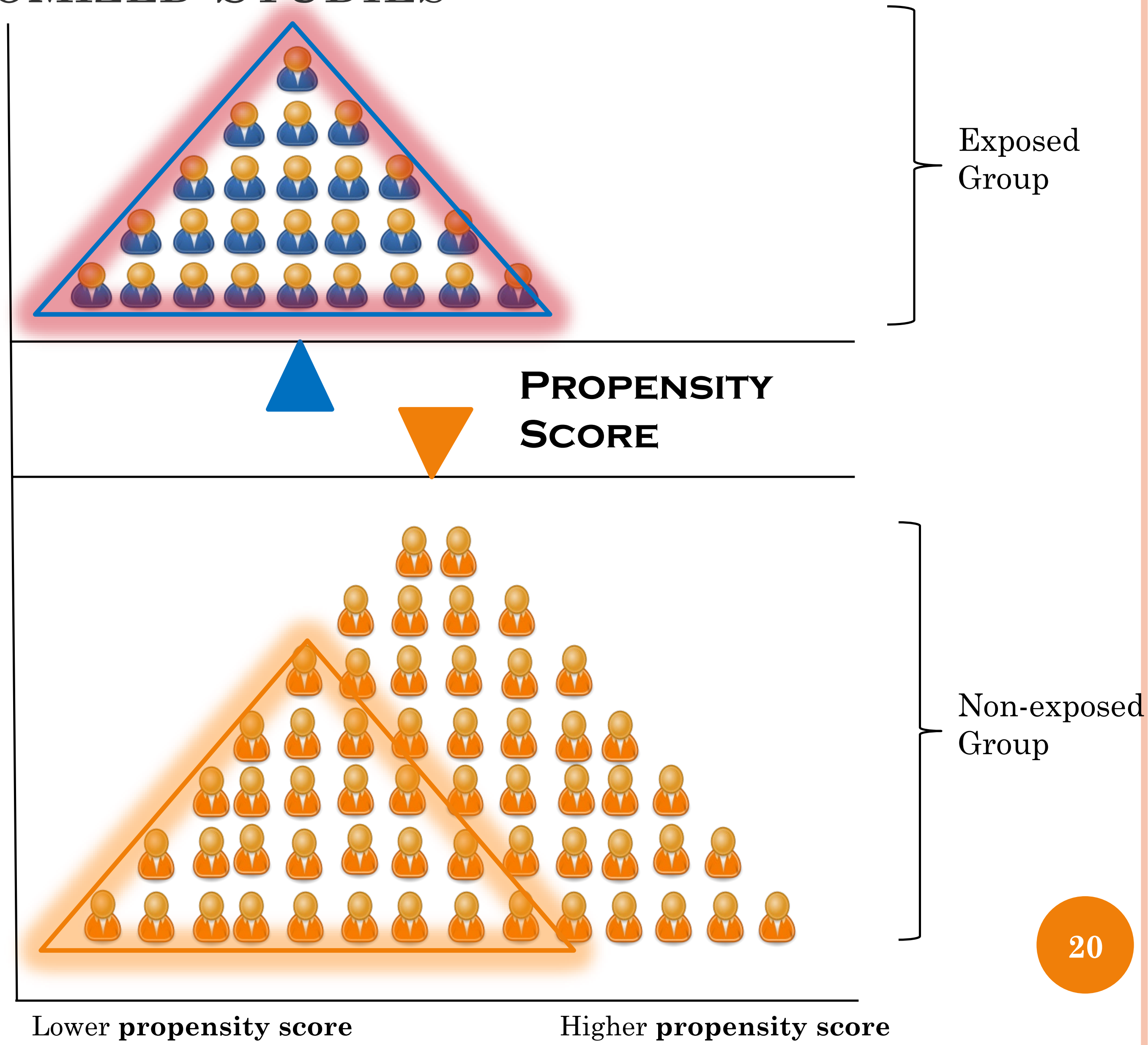
The Need for Propensity Score Matching

- ***Multiple*** baseline characteristics
  - Propensity scores reduce all these characteristics to a *single measure per individual*
    - **This single measure is the probability of exposure given the baseline characteristics**
  - Individuals are matched by propensity score, rather than by each individual baseline characteristic
    - Each individual in the exposure group is matched with an individual having a **similar propensity score** in the non-exposure group
    - For each individual, we can observe the outcome after exposure or outcome after non-exposure, but we cannot observe both
    - The 'matched' non-exposed individuals will assist in estimating the effect of non-exposure for the 'paired' exposed subjects

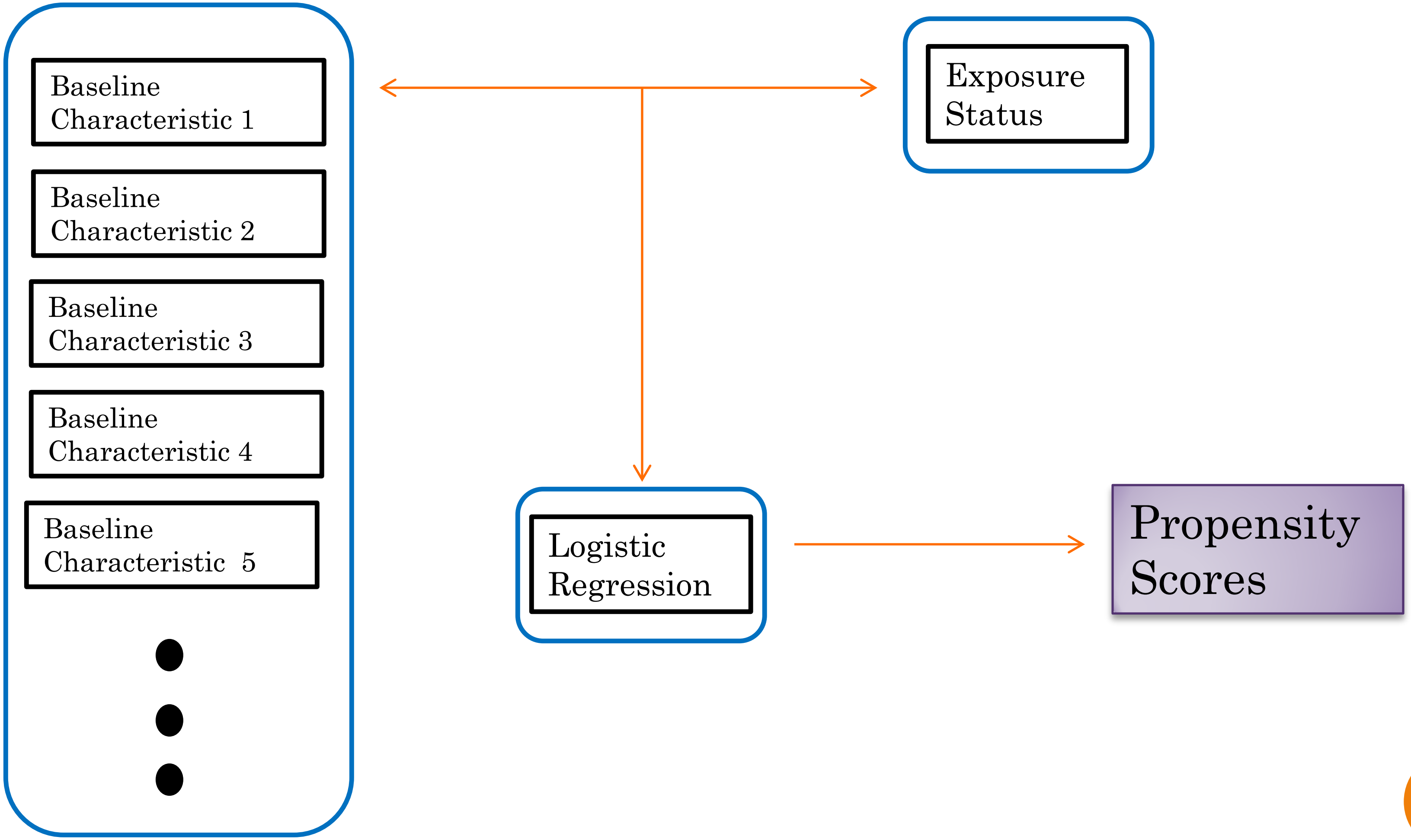
Non-Randomized Studies

When multiple baseline characteristics are present, we match by propensity scores.

Well-known that matching by propensity scores reduces biases (Rosenbaum & Rubin 1985)



Propensity Scores Building Blocks


Baseline Characteristic 1

Baseline Characteristic 2

Baseline Characteristic 3

Baseline Characteristic 4

Baseline Characteristic 5

Exposure Status

Logistic Regression

Propensity Scores

Propensity Score Properties

- Propensity Score (PS) → ***predicted probability of exposure/intervention/treatment given observed baseline characteristics***
- PS is a ***balancing score***
  - Conditional on the PS, the distribution of observed baseline factors is similar between exposed & unexposed individuals
- PS is (usually) estimated with logistic regression
  - Exposure status regressed on observed baseline covariates
  - Other methods also used (see Austin 2011)
- PS ***reduces*** average treatment effect estimation ***biases* in observational studies**

Propensity Score Matching Summary

- **PSM**: process of forming a matched set of exposed & unexposed subjects with a similar PS to estimate the average treatment effect

- Using a matched sample, direct comparisons of outcomes between exposed & unexposed groups are made **to estimate average treatment effects with reduced bias**

- **Emulates analysis of treatment effects in RCTs**

OUTLINE

- Medial/Population Health Motivation
- Study Design Issues
- Purpose of Propensity Score Matching (PSM)
- **Implementation of PSM & Balance Diagnostics**
- Application to Treatment in Crohn's Disease Using CPRD Data
- Next Steps/Other Areas of Application

Techniques In Forming A Matched Sample: I

- **Matching With vs. Without Replacement**
  - With: an unexposed subject can be matched with multiple exposed individuals
  - Without: an unexposed subject is matched with only 1 exposed person

Techniques In Forming A Matched Sample: II

- **Greedy vs. Optimal Matching**
  - **<u>Greedy</u>**
    - Exposed subject selected at random
    - Unexposed subject with closest PS to that of the randomly selected exposed subject is chosen for matching
      - Nearest neighbor matching
      - Nearest neighbor within a pre-specified ***caliper distance***
        - Restricted so that absolute difference in PSs is within threshold
        - If no unexposed subjects meet threshold, then exposed subject is not matched & is discarded
    - At each step, the nearest unexposed subject is chosen to be matched with the exposed subject, even if the unexposed subject has a PS that would better match a different exposed subject
    - **Sequential process until all exposed subjects are matched**

Techniques In Forming A Matched Sample: III

- **Optimal Matching**
  - Subjects are matched to minimize a global distance measure
  - Smallest average absolute distance across all within-pair differences of the PS
  - Greedy & Optimal yield *similar* results (Gu & Rosenbaum 1993)
  - However, Optimal matching is better at minimizing within-pair differences & is preferred when there are fewer control matches for the exposed subjects
  - Greedy is faster, but Optimal is more robust (Gu & Rosenbaum 1993)

Matching Ratios

- 1:1 (pair) matching (most common)
- M:1 (many to one) matching
  - M unexposed subjects matched to a single exposed subject
  - Choice of M is both a science & an art
    - M too small may discard too many unexposed samples (inefficient)
    - M too large may lead to biased samples (PSs may be too dissimilar between groups)

- Matched sets of either:
    - 1 exposed subject to at least 1 unexposed
    - 1 unexposed subject to at least 1 exposed

Balance Diagnostics

- Is the PS model specified appropriately?
- Within the PSM sample:
  - **Are distributions of measured baseline characteristics *similar* between exposed & unexposed subjects with similar PSs?**
  - **Numerical** methods: Are the standardized mean (or prevalence) differences between exposed & unexposed subjects for the covariates small?
    - Although there is no global agreement on threshold for 'small enough', 0.1 units is widely accepted (Normand et al. 2001)
  - **Graphical** methods: boxplots, Q-Q plots, or CDFs

Results of Balance Diagnostics

- If there are still differences, then the model may need modifying
  - Add more covariates
  - Include covariate interactions
- Continue process of modifications & balance-checking until differences are negligible

PSM vs. Regression Adjustment

- For continuous outcomes modeled through linear regression
  - PSM & regression adjustments yield more similar results (Rosenbaum 2005)

- For binary, multi-category, & time-to-event outcomes
  - PSM yields ORs & HRs that reduce bias vs. regression (Austin et al. 2007)

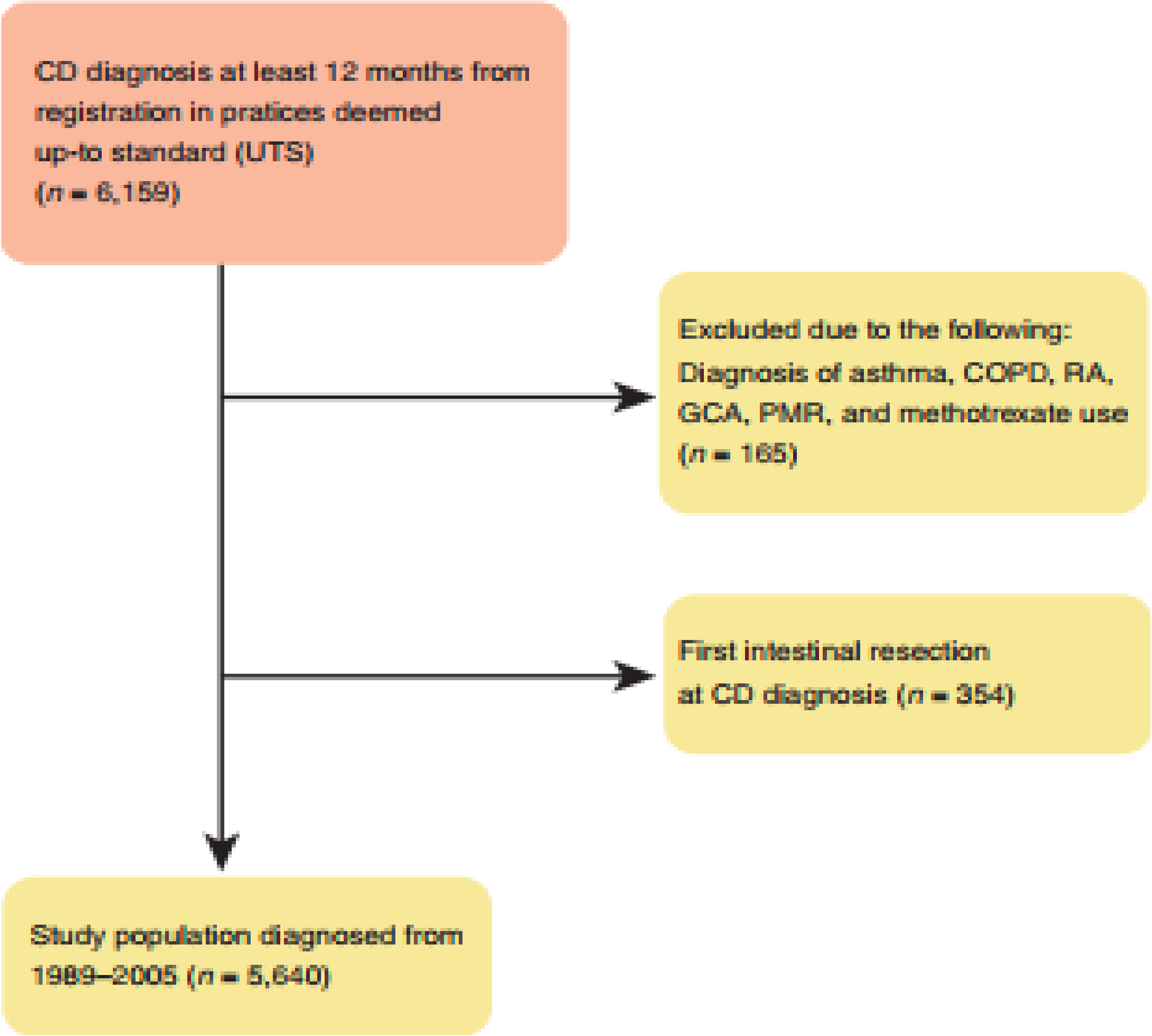
Application in R Statistical Software

- *MatchIt/Matching/PSAgraphics/…*
  - R packages
  - Prepares the observational data for balanced exposure & non-exposure groups prior to parametric analysis (e.g., survival analysis, etc.)
  - Calculate average treatment effect with reduced bias
  - Specify matching method (Greedy/Optimal, Ratio Type, etc.)
  - Balance diagnostics & graphical displays

Outline

- Medical/Population Health Motivation
- Study Design Issues
- Purpose of Propensity Score Matching (PSM)
- Implementation of PSM & Balance Diagnostics
- **Application to Treatment in Crohn's Disease Using CPRD Data**
- Next Steps/Other Areas of Application

Flow Diagram for UK CPRD Crohn's Cohort Study



CD diagnosis at least 12 months from registration in pratices deemed up-to standard (UTS) (*n* = 6,159)

Excluded due to the following: Diagnosis of asthma, COPD, RA, GCA, PMR, and methotrexate use (*n* = 165)

First intestinal resection at CD diagnosis (*n* = 354)

Study population diagnosed from 1989–2005 (*n* = 5,640)

THIOPURINE (TP) PRESCRIPTION

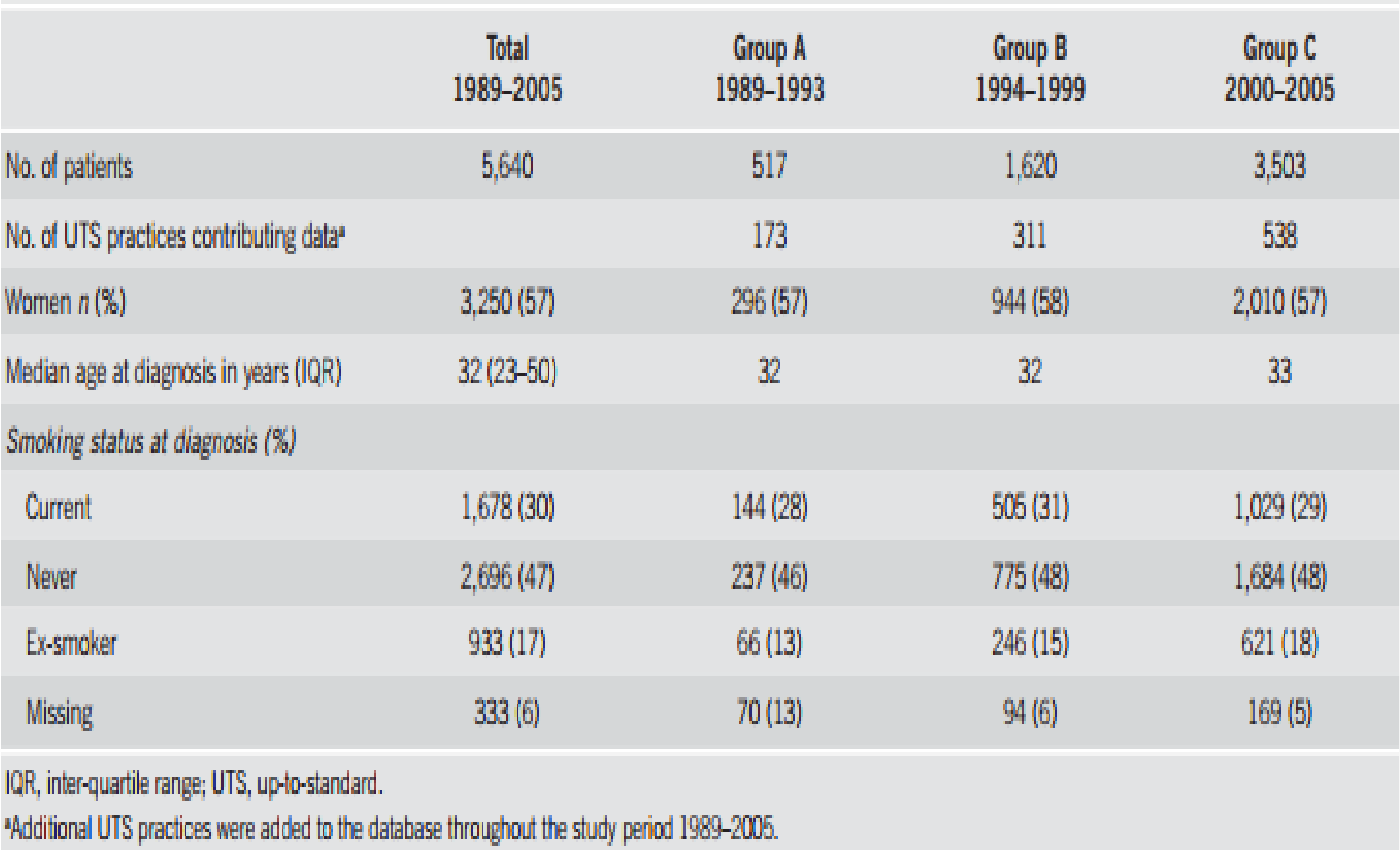
- TP ‘users’
  - ≥ 1 prescription before surgery (or during follow-up, if no surgery)
  - 25% users
- TP ‘non-users’
  - No prescription or 1st prescription after 1st surgical resection
- Early Use
  - Initiated TP within 1st year of diagnosis
- Late Use
  - Initiation after 1st year

- Trends in TP prescribing & first resection
- Compare patients treated with vs. without TPs for:
  - First resection rates
  - Does early or prolonged use impact surgery risk?

DATA

- Treatment duration
  - $\geq$ 6 months
  - $\geq$ 12 months
- Primary Outcome: 1st intestinal resection
- Potential Confounders
  - age of diagnosis; gender; year of diagnosis; history of appendectomy; smoking; 5-aminosalicylic acid (5-ASA); corticosteroid
- Missing Data
  - Due to high level of completeness, used a complete case analysis, thus excluding patients with missing information (~5% of data)

Patient Characteristics at CD Diagnosis



	Total 1989–2005	Group A 1989–1993	Group B 1994–1999	Group C 2000–2005
No. of patients	5,640	517	1,620	3,503
No. of UTS practices contributing data[a]		173	311	538
Women *n* (%)	3,250 (57)	296 (57)	944 (58)	2,010 (57)
Median age at diagnosis in years (IQR)	32 (23–50)	32	32	33
Smoking status at diagnosis (%)				
Current	1,678 (30)	144 (28)	505 (31)	1,029 (29)
Never	2,696 (47)	237 (46)	775 (48)	1,684 (48)
Ex-smoker	933 (17)	66 (13)	246 (15)	621 (18)
Missing	333 (6)	70 (13)	94 (6)	169 (5)

IQR, inter-quartile range; UTS, up-to-standard.

[a] Additional UTS practices were added to the database throughout the study period 1989–2005.

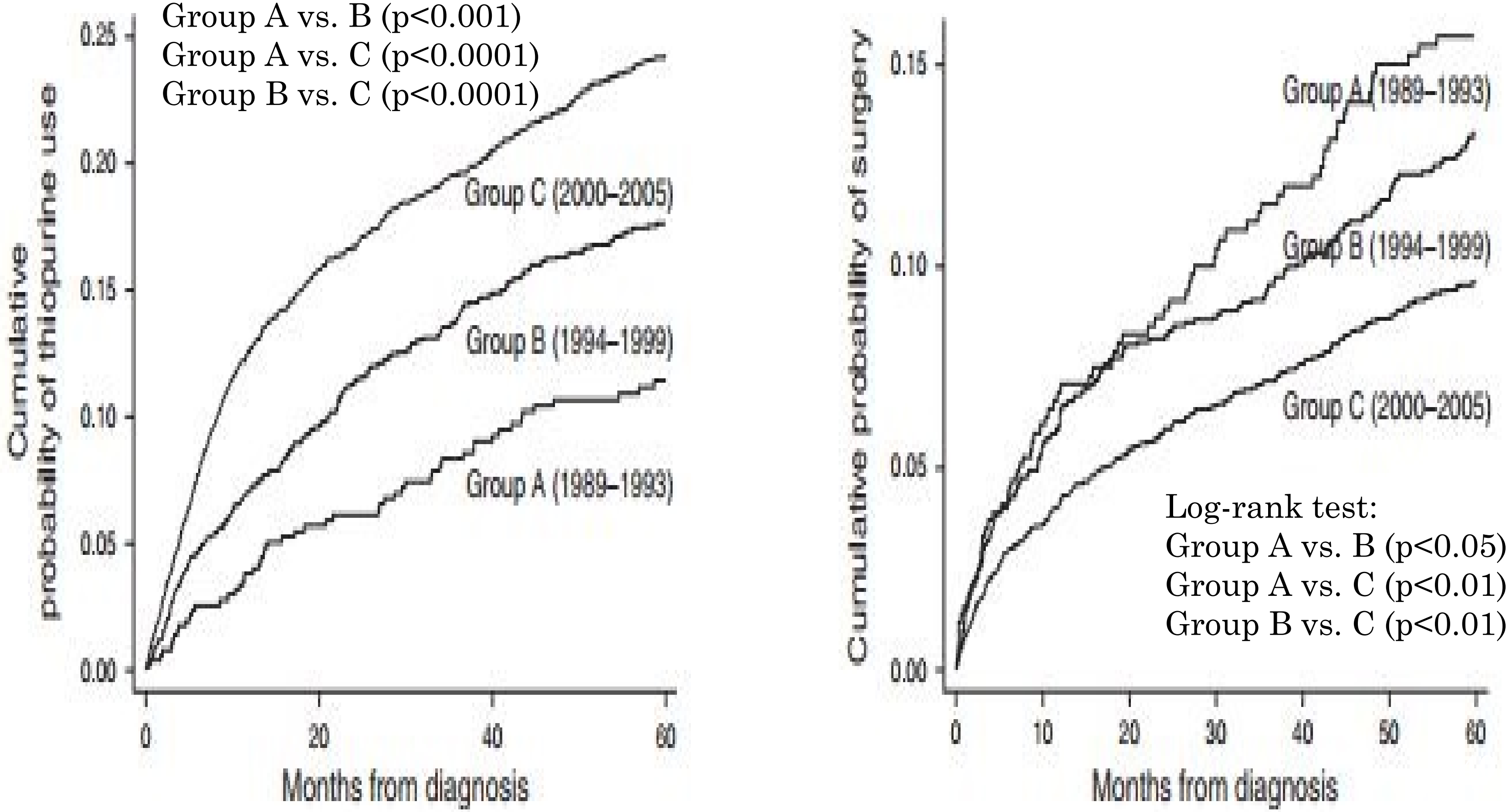
Propensity Score Matching As a Remedy

- To account for inherent selection bias that exists from a historical cohort study
- Calculate PS of thiopurine treatment allocation
- Included all patient-specific covariates into a multivariate logistic regression model to compute PSs for exposed (TP users) & unexposed (non-users)
- 2:1 Optimal matching
- Checked balance diagnostics
- Sensitivity analyses
  - 3:1; 2:1; & 1:1 matching ratios
  - Greedy matching

Cox Proportional Hazards Model

- Applied Cox-PH on **matched data** to determine effect of TP use on 1st intestinal resection in CD patients
- Hazard Ratios (HRs) & 95% CIs
- Sensitivity analysis
  - Greedy matching: *similar* results, however 5-ASA becomes non-significant (p=0.1223) on risk of surgery
    - Nearly 80% of the matched controls were the same in both algorithms
  - Balance diagnostics
    - 1:1 Optimal matching provided best balance diagnostics with all covariate mean & prevalence differences within 0.04 units, however this ratio discards the most data
    - 2:1 Optimal: all differences within 0.06 units except 1 covariate within 0.1 units
    - 3:1 Optimal: all within 0.09 units except 1 covariate within 0.2 units

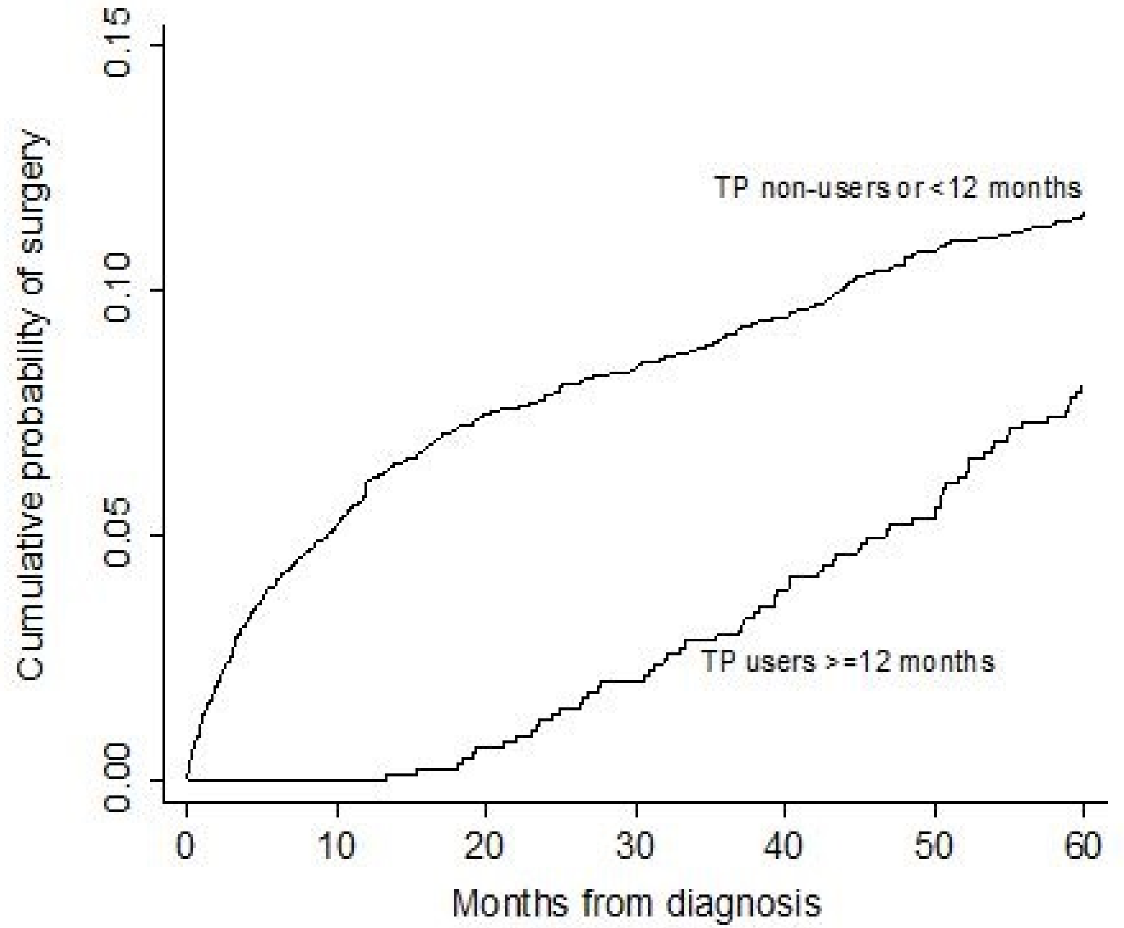
Kaplan- Meier Curve Comparing Cumulative Probability of: TP Use (left) & 1st Intestinal Surgery (right) after CD Diagnosis by Period of Diagnosis



Log-rank test:
Group A vs. B ($p<0.001$)
Group A vs. C ($p<0.0001$)
Group B vs. C ($p<0.0001$)

Log-rank test:
Group A vs. B ($p<0.05$)
Group A vs. C ($p<0.01$)
Group B vs. C ($p<0.01$)

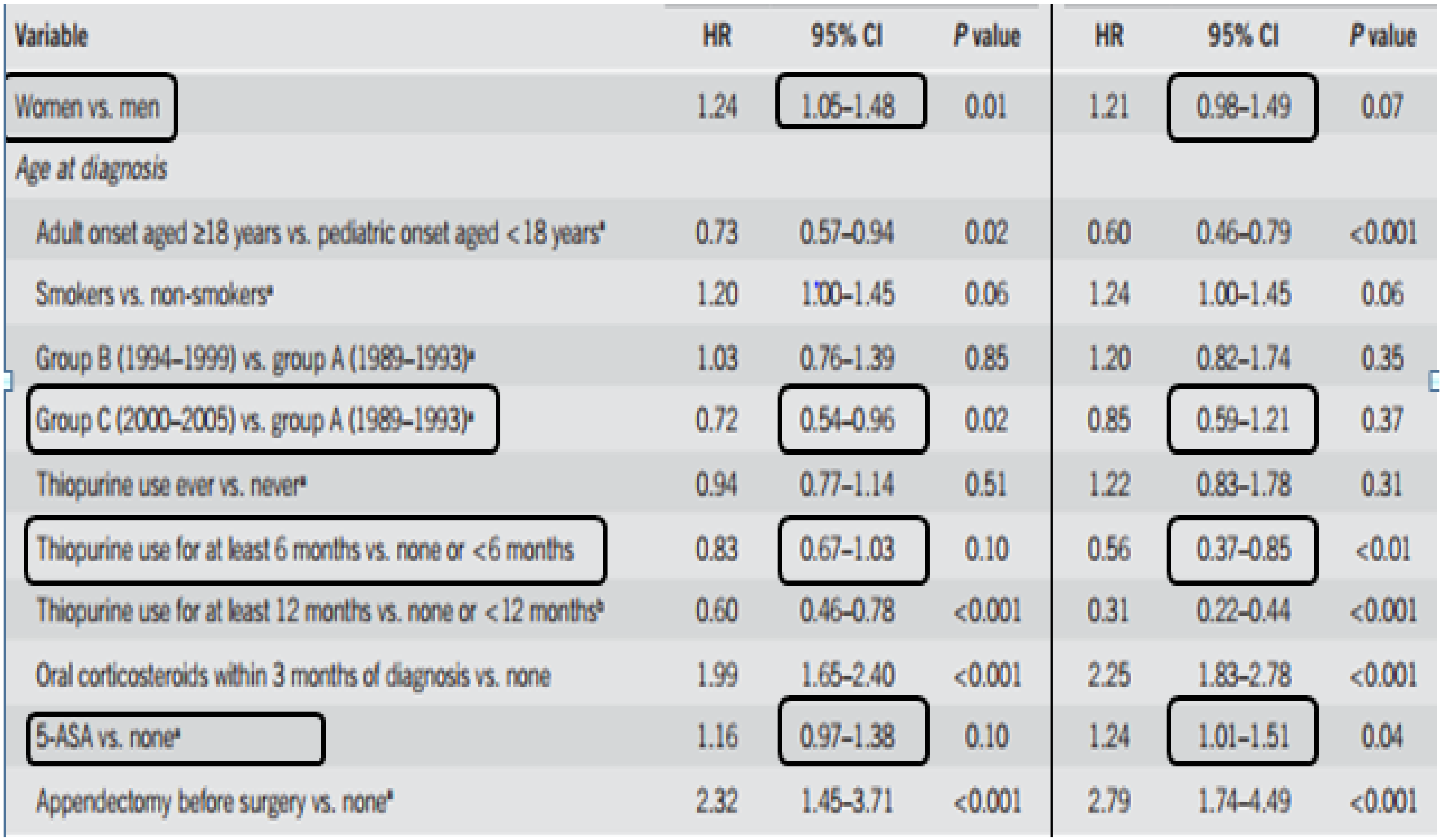
K-M Curve Comparing 5-Year Cumulative Probability of Surgery in CD Patients Receiving ≥12 Months of Thiopurine (TP) vs. Non-Users or Those Who Receiving <12 Months of therapy



Log-rank test:
p<0.001

What are Risk & Protective Factors for Surgery?

Cox Regression after Adjusting for 2:1 Optimal Propensity Score Matching (right), Showing Hazard Ratios (HRs) for Risk of Surgery within 5 Years of CD Diagnosis



	Before PSM			After PSM		
Variable	HR	95% CI	*P* value	HR	95% CI	*P* value
Women vs. men	1.24	1.05–1.48	0.01	1.21	0.98–1.49	0.07
Age at diagnosis						
Adult onset aged ≥18 years vs. pediatric onset aged <18 years[a]	0.73	0.57–0.94	0.02	0.60	0.46–0.79	<0.001
Smokers vs. non-smokers[a]	1.20	1.00–1.45	0.06	1.24	1.00–1.45	0.06
Group B (1994–1999) vs. group A (1989–1993)[a]	1.03	0.76–1.39	0.85	1.20	0.82–1.74	0.35
Group C (2000–2005) vs. group A (1989–1993)[a]	0.72	0.54–0.96	0.02	0.85	0.59–1.21	0.37
Thiopurine use ever vs. never[a]	0.94	0.77–1.14	0.51	1.22	0.83–1.78	0.31
Thiopurine use for at least 6 months vs. none or <6 months	0.83	0.67–1.03	0.10	0.56	0.37–0.85	<0.01
Thiopurine use for at least 12 months vs. none or <12 months[a]	0.60	0.46–0.78	<0.001	0.31	0.22–0.44	<0.001
Oral corticosteroids within 3 months of diagnosis vs. none	1.99	1.65–2.40	<0.001	2.25	1.83–2.78	<0.001
5-ASA vs. none[a]	1.16	0.97–1.38	0.10	1.24	1.01–1.51	0.04
Appendectomy before surgery vs. none[a]	2.32	1.45–3.71	<0.001	2.79	1.74–4.49	<0.001

[1]Reference group. Analysis based on 2:1 optimal matching between TP users and non-users (n=3,693),

*All multivariate results are shown for the model including TP use for ≥12 months (omitting 6 months duration). TP use for ≥6 months was added separately when the 12 months duration was removed; this was due to substantial multicollinearity present in the model.

- *When should therapy be initiated?*
  - *Sub-analysis of 879 CD patients with ≥ 12 months of therapy showed that both early (within 1 yr of diagnosis) & late (after 1 yr of diagnosis) initiation reduced risk of surgery*
  - *Early: HR=0.41; 95% CI: (0.27,0.61)*
  - *Late: HR=0.21; 95% CI: (0.13,0.34)*

More Application Details Available In:

- The Impact of Timing and Duration of Thiopurine Treatment on First Intestinal Resection in Crohn's Disease: National UK Population-Based Study 1989-2010.
- *American Journal of Gastroenterology*. 2014;109:409-16. https://www.nature.com/articles/ajg2013462
- Joint work with:
  - Sukh Chatu
    - Consultant Gastroenterologist & Physician, King's College Hospital, London
  - Richard Pollok
    - Dept. of Gastroenterology, St. George's University Hospital, London
  - Azeem Majeed
    - Professor & Dept. Head, Primary Care & Public Health, Imperial College London
  - Sonia Saxena
    - Professor, Primary Care & Public Health, Child Health Unit, Imperial College London
  - Ghasem Yadegarfar
    - Dept. of Primary Care and Public Health, Imperial College London
  - Venkat Subramanian
    - Clinical Associate Professor & Honorary Consultant Gastroenterologist, Leeds Institute of Biomedical & Clinical Sciences, University of Leeds;
      Dept. of Gastroenterology, St. James's University Hospital, Leeds
  - Vasa Curcin
    - Dept. of Computing, Imperial College London

Outline

- Medical/Population Health Motivation
- Study Design Issues
- Purpose of Propensity Score Matching (PSM)
- Implementation of PSM & Balance Diagnostics
- Application to Treatment in Crohn's Disease Using CPRD Data
- **Next Steps/Other Areas of Application**

NEXT STEPS: BIG DATA & STATISTICS

- Multi-year cohort study on U.K. diabetic population to assess effect of exposure status (meeting QOF/NDA targets) on hospital admissions & mortality
  - Propensity score matching methods needed
- U.K. Population data with tens of millions of observations
  - CPRD diabetic population (2010-2017)
  - Joint work with colleagues at School of Public Health, Faculty of Medicine, Imperial College London
  - Big Data & Analytical Unit, Imperial College London
- New wave of data management & methods
  - Statistical methods to handle increasingly large data
  - Messy data from multiple sources require combining & cleaning

THANK YOU

- **Questions?**

- **References**

- Austin, PC, Grootendorst P, Normand SL, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. Statistics in Medicine 2007;26(4):754-68.
- Austin P. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research 2011;46:399-424.
- Cosnes J, Cattan S, Blain A *et al*. Long-term evolution of disease behavior of Crohn's disease. Inflamm Bowel Dis 2002; 8:244-50.
- Gu XS, Rosenbaum, PR. Comparison of multivariate matching methods: structures, distances, and algorithms. Journal of Computational and Graphical Statistics 1993;2:405-20.
- Ho DE, Imai K, King G, Stuart EA. MatchIt: Nonparametric preprocessing for parametric causal inference. Journal of Statistical Software 2011;42:1-28.
- Lewis JD, Brensinger C, Bilker WB *et al.* Validity and completeness of the General Practice Research Database for studies of inflammatory bowel disease . Pharmacoepidemiol Drug Saf 2002;11:211-8.
- Lewis JD, Aberra FN, Lichtenstein GR *et al.* Seasonal variation in flares of inflammatory bowel disease. Gastroenterology 2004;126:665-73.
- Loftus EV Jr. Crohn's disease: why the disparity in mortality? Gut 2006;55:447-9.
- Normand T, Landrum MB, Guadagnoli E *et al.* Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: a matched analysis using propensity scores. Journal of Clinical Epidemiology 2001;54:387–398.
- Prefontaine E, Sutherland LR, MacDonald JK *et al.* Azathioprine or 6-mercaptopurine for maintenance of remission in Crohn's disease . Cochrane Database Syst Rev 2009; 9:CD000067.
- Ramadas AV, Gunesh S, Thomas GAO *et al.* Natural history of Crohn's disease in a population-based cohort from Cardiff (1986-2003): a study of changes in medical treatment and surgical resection rates. Gut 2010; 59:1200-6.
- Rosenbaum PR. Propensity score. In: Armitage P and Colton T. editors. Encyclopedia of biostatistics. 2nd ed. Boston, MA: Wiley;2005.4267-72.
- Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician 1985;39:33–8.