Identifying Causal Effects from Electronic Health Record Data

Victor M. Castro

March 1, 2018

MASSACHUSETTS GENERAL HOSPITAL

PARTNERS® HEALTHCARE

Introduction

- **Partners Healthcare**
  - 1.2 million patients seen annually in Eastern Massachusetts
  - $1.5B Research Enterprise
  - 68k clinicians and staff

Founding Members

- Brigham and Women's Hospital
- Massachusetts General Hospital

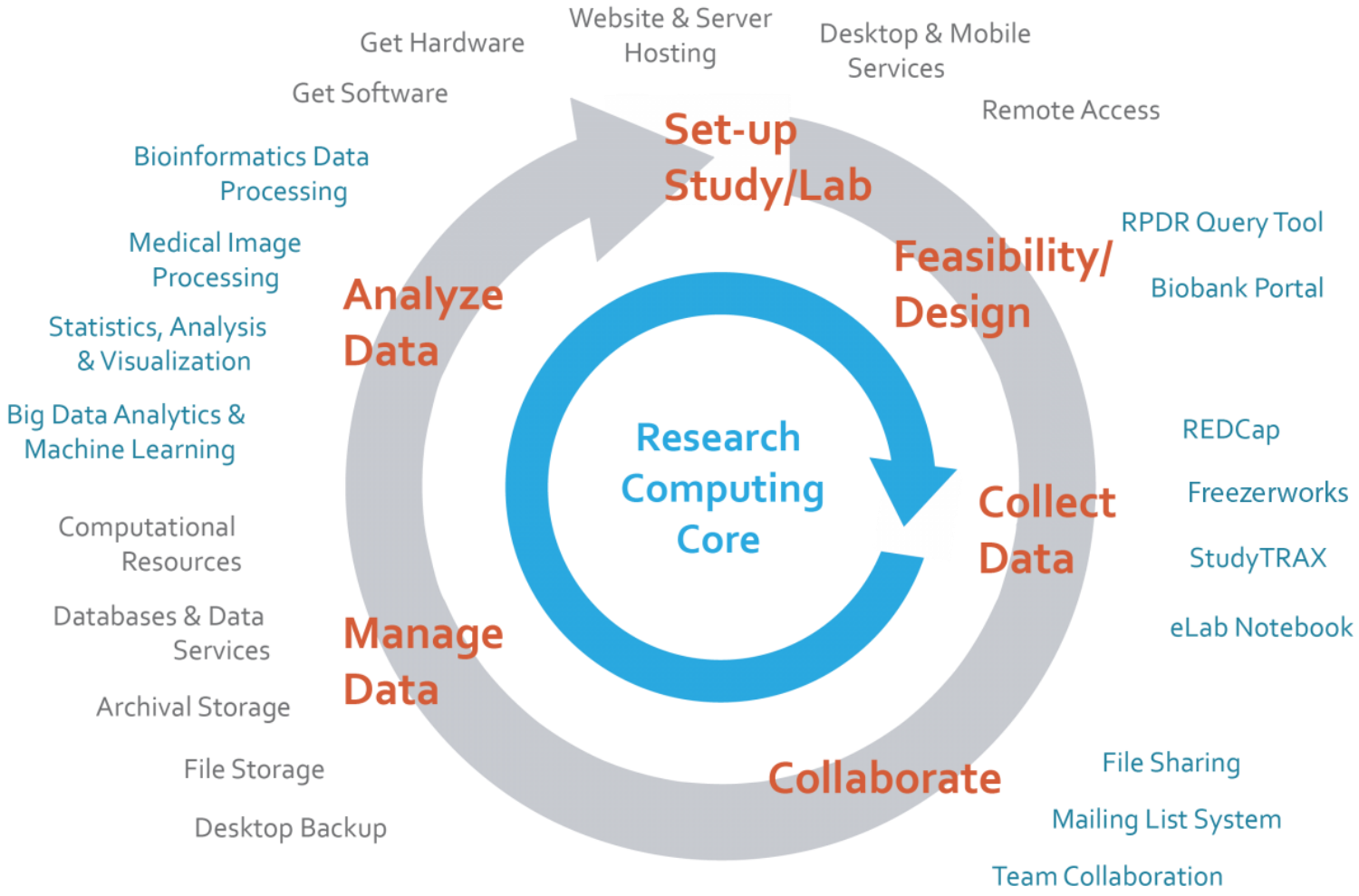
Members

- Brigham and Women's Faulkner Hospital
- Cooley Dickinson Hospital
- Martha's Vineyard Hospital
- McLean Hospital
- MGH Institute of Health Professions
- Nantucket Cottage Hospital
- Neighborhood Health Plan
- Newton-Wellesley Hospital
- North Shore Medical Center
- Partners Community Physicians Organization
- Partners HealthCare at Home
- Spaulding Rehabilitation Network
- Wentworth-Douglass Hospital

PARTNERS HEALTHCARE

RISC RESEARCH INFORMATION SCIENCE & COMPUTING

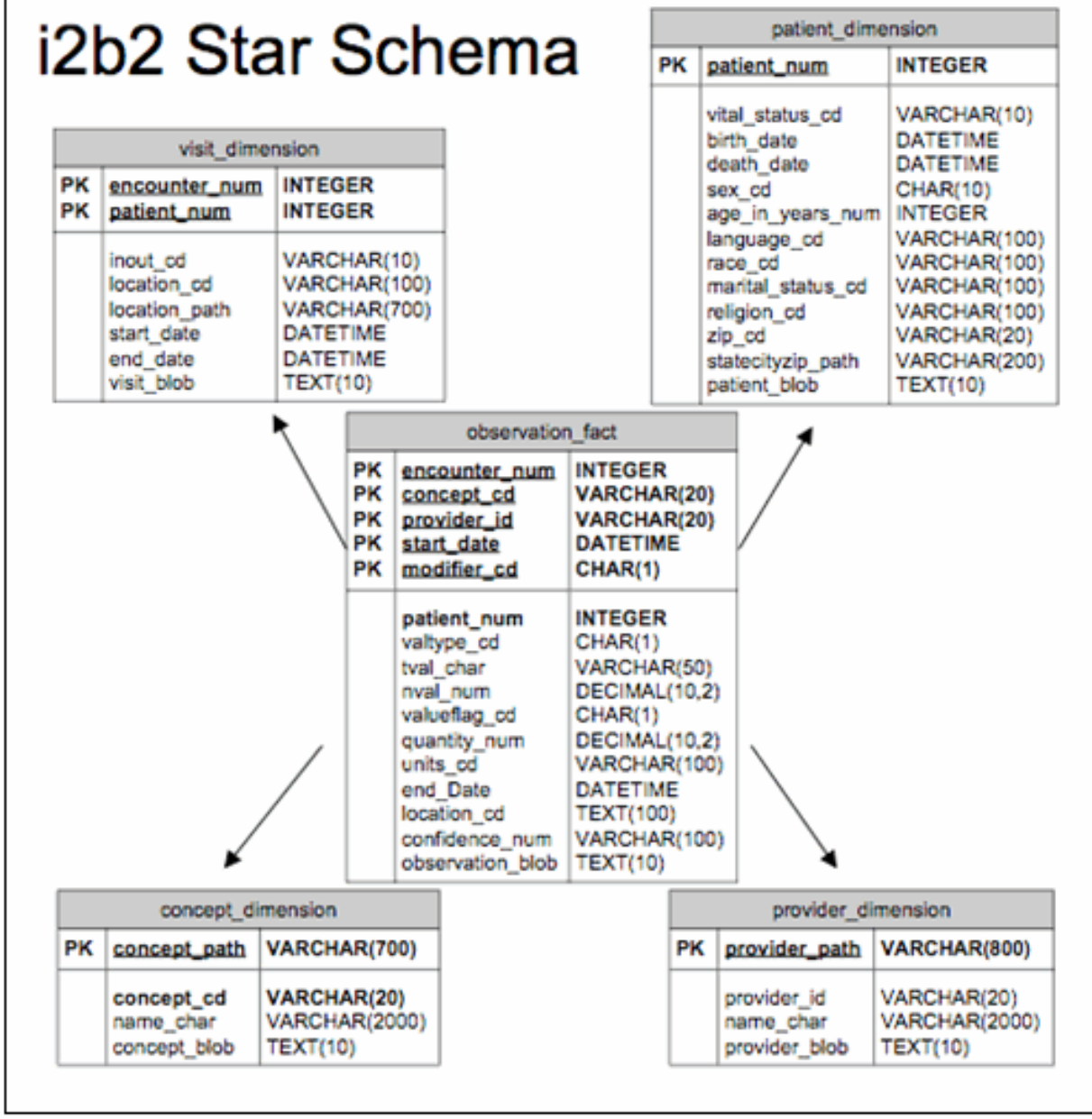
We support research by providing scientific services and technology, a centralized clinical data registry, genomics IT, specimen banking, and administrative systems



Research Patient Data Registry

Patients	**4.2 million**
Diagnosis	360 million
Medications	318 million
Procedures	542 million
Vital Signs	199 million
Lab tests	1.2 billion
Clinical Notes	170 million
TOTAL FACTS	3 billion

i2b2 Star Schema



Phenotype Discovery Center

- The **Partners Phenotype Discovery Center (PDC)** is developing computational methods and platforms to help harness the power of big data in the field of medical research across all Partners HealthCare institutions

- Identify true causal associations from EHR data (primarily from the RPDR)
  - Includes associations between EHR-derived phenotypes and genotypes and environmental variables

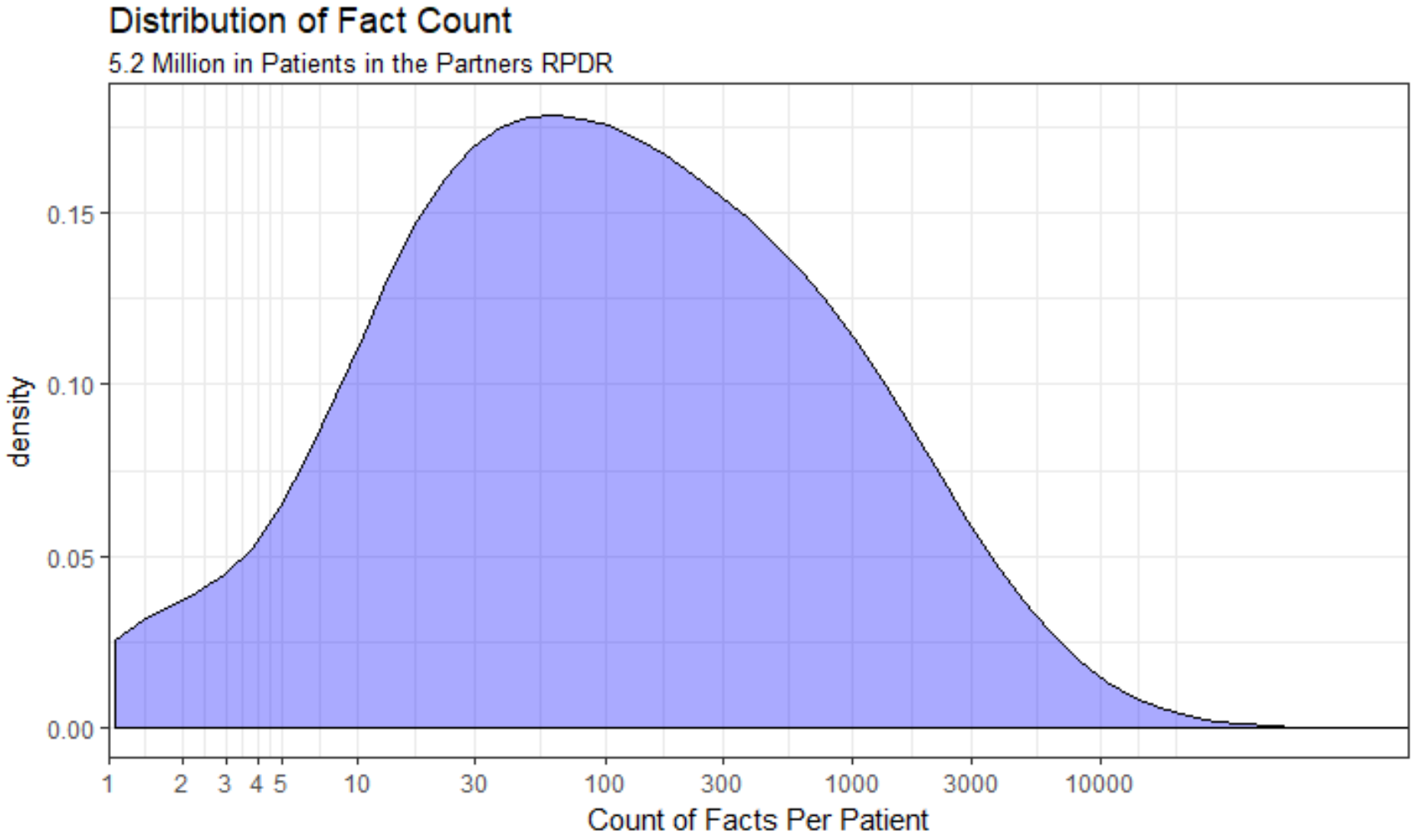
Confounding in EHR Data

- Data quality / inaccuracy
- Confounding by unmeasured observations (the open system problem)
- Onset of disease may not be well documented
-
-
-
- **Confounding by utilization | data completeness**

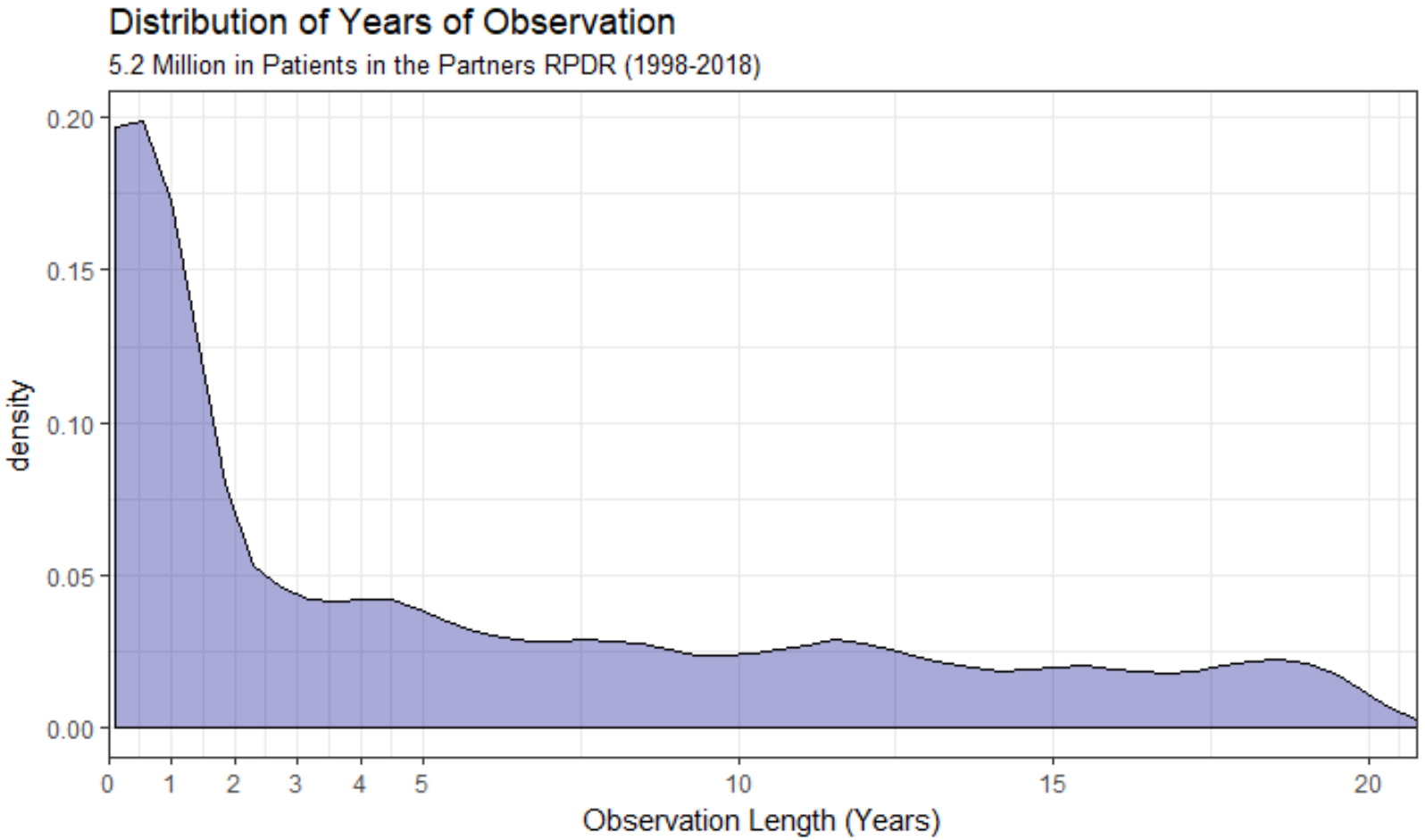
Measuring Utilization

- Count of “Facts”
- Count of Visits (Inpatient, Outpatient, ER)
- Count of Notes
- Count of Diagnosis
- Count of Procedures
- Comorbidity Indexes (Charlson)

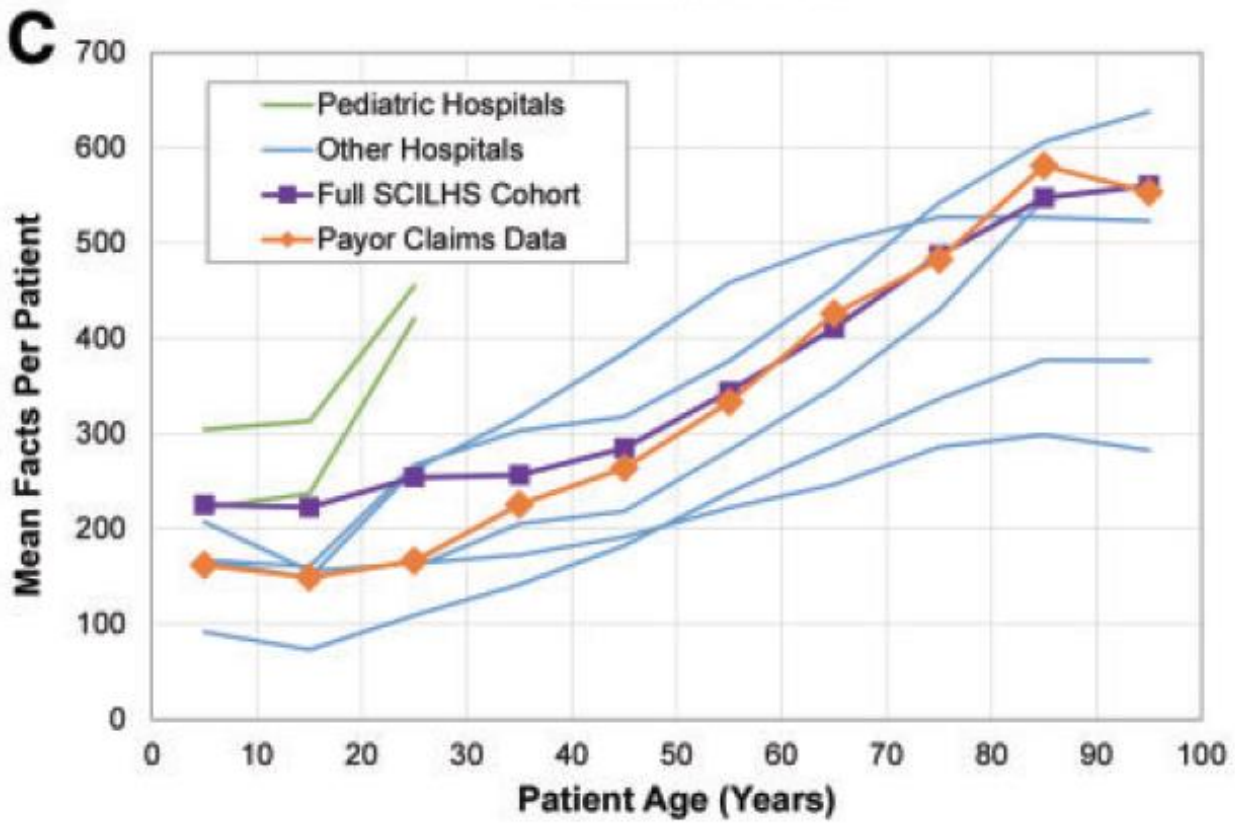
Huge variation in utilization



Variation in Observation Length



Utilization by Age



Webber, et al. JAMIA 2017

Data Completeness

- Observation length
- Fact quartile
- Presence of study-related data types
  - E.g. for a pharmacovigilance study, require population to have at least 1 diagnosis and 1 medication
- Data Floor <> Minimum number of visits
- Loyalty cohort

Biases introduced by filtering electronic health records for patients with "complete data"

Griffin M Weber,[1,2] William G Adams,[3] Elmer V Bernstam,[4] Jonathan P Bickel,[5] Kathe P Fox,[6] Keith Marsolo,[7] Vijay A Raghavan,[8] Alexander Turchin,[9] Xiaobo Zhou,[10] Shawn N Murphy,[11] and Kenneth D Mandl[1,5]

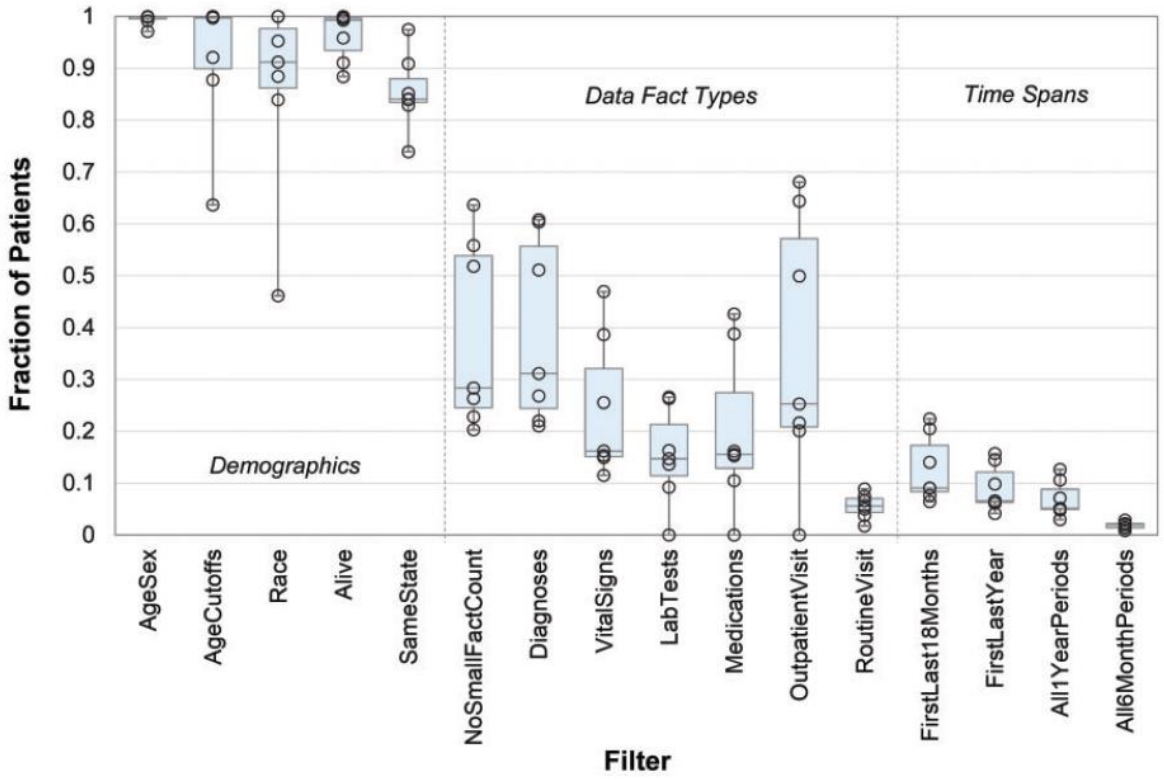
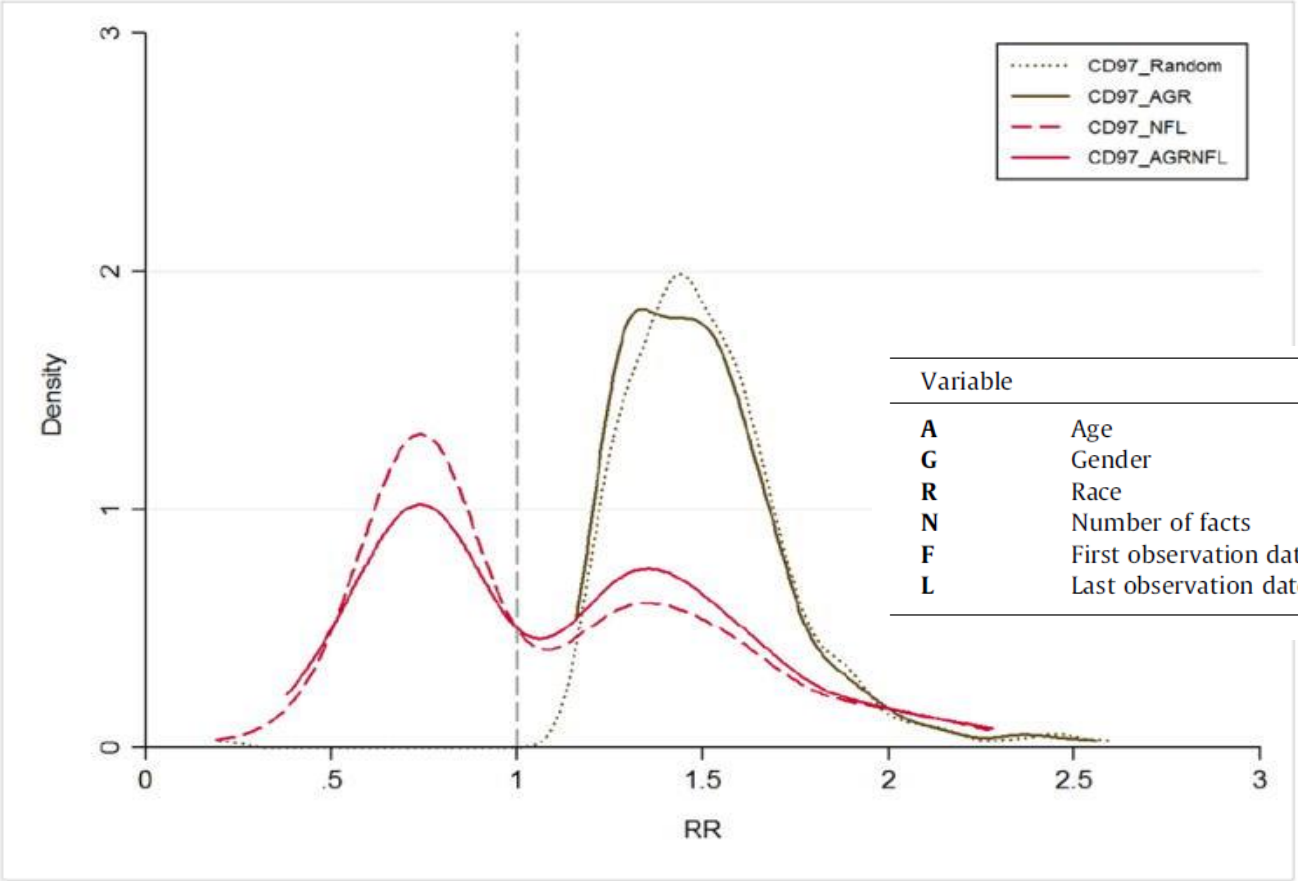


Figure 2. Fraction of patients at each of the 7 electronic health record sites who passed each of the 16 filters. Boxes indicate the median and quartiles.

Matching to Minimize Confounding

Evaluation of matched control algorithms in EHR-based phenotyping studies: A case study of inflammatory bowel disease comorbidities ☆

Victor M. Castro [a,*], W. Kay Apperson [a], Vivian S. Gainer [a], Ashwin N. Ananthakrishnan [c], Alyssa P. Goodson [a], Taowei D. Wang [a], Christopher D. Herrick [a], Shawn N. Murphy [a,b]



Variable	
A	Age
G	Gender
R	Race
N	Number of facts
F	First observation date
L	Last observation date

Table 3a
Selected comorbidity associations with CD.

Comorbidity (lifetime history)	CD vs. random		CD vs. AGR		CD vs. NFL		CD vs. AGRNFL	
	RR		RR		RR		RR	
Asthma	1.30	+	1.28	+	0.88		0.90	
Chronic airway obstruction	1.34	+	1.39	+	0.86		1.02	
Chronic kidney disease	1.55	+	1.53	+	1.05		1.15	+
Congenital anomalies	1.29	+	1.35	+	0.97		0.95	
Diabetes mellitus, Type 2	1.13		1.26	+	0.76	–	0.87	
Disorders of lipoid metabolism	1.19	+	1.17	+	0.68	–	0.78	–
Fractures	1.09		1.06		0.78	–	0.72	–
Gastrointestinal hemorrhage	1.81	+	1.86	+	1.42	+	1.48	+
Headaches	1.27	+	1.21	+	0.82	–	0.80	–
Heart failure	1.26	+	1.31	+	0.80	–	0.91	
Hypertension	1.23	+	1.25	+	0.76	–	0.89	–
Ischemic heart disease	1.17	+	1.24	+	0.76	–	0.90	
Major depression	1.47	+	1.38	+	0.94		0.91	
Malignant neoplasm	1.11		1.10		0.82	–	0.83	–
Osteoarthrosis	1.28	+	1.19	+	0.78	–	0.87	–
Rheumatoid arthritis	1.59	+	1.69	+	1.40	+	1.41	+

RR: relative risk.
+: significant positive association (RR > 1).
–: significant negative association (RR < 1).

Matching Challenges

- Can be difficult to find controls in heterogenous data, even with large sample sizes

- Many matching methods improve balance on 1 variable while reducing it on others

Matching Methods

- Debate on optimal matching methods:
  - Propensity score matching
  - Genetic matching
  - Coarsened Exact Matching (CEM)
  - Others

- We focus on CEM because:
  - Easy to implement
  - Transparency of design choices
  - Incorporation of heuristics
  - Stable

Coarsened Exact Matching

Volume 20, Issue 1 Winter 2012 , pp. 1-24

Causal Inference without Balance Checking: Coarsened Exact Matching

Stefano M. Iacus [(a1)], Gary King [(a2)] and Giuseppe Porro [(a3)]

https://doi.org/10.1093/pan/mpr013 Published online: 04 January 2017

- Coarsen matching variables based on heuristics or common histogram binning techniques
- Perform exact matching on the coarsened variables
- Group observations into strata
- Prune any stratum with 0 case or 0 controls
- “Sacrifice some data to avoid bias” -Blackwell

Coarsened Exact Matching

Examples of Coarsened Variables

Age at index = 38.5 years	*coarsened to*	35-39
Year of index event = 2009	*coarsened to*	2009-2012
Gender = F	*not coarsened*	F
Race = Asian	*coarsened to*	Non-Caucasian

Implemented using cem R package

Ri2b2matchcontrols

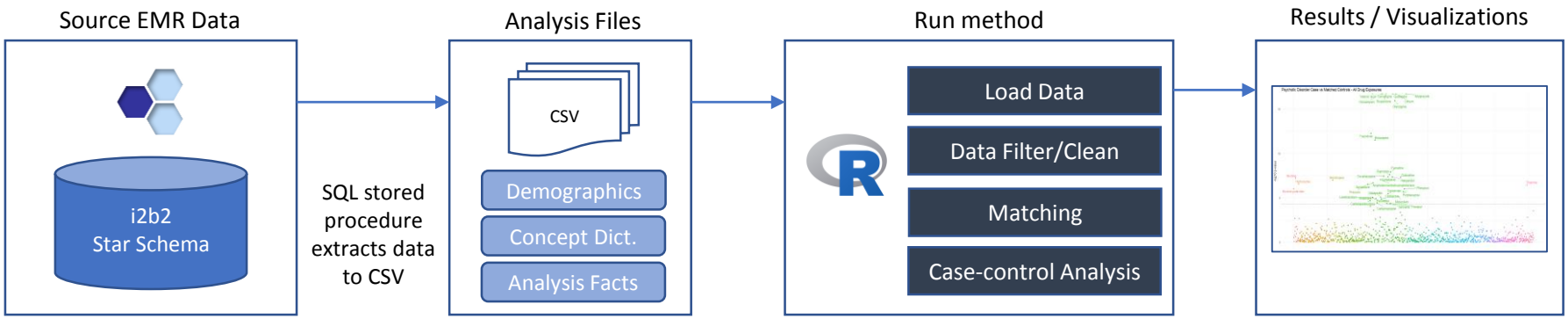
- Build a tool to identify causal drug effects in large EHR datasets
- Integrate with i2b2 framework to enable reproducibility and generalizability across many sites
- Control for confounding typically encountered in observation EHR data:
  - Confounding by unmeasured observations (the open system problem)
  - Confounding by severity of disease/utilization:
    - Sicker patients visit the doctor/get admitted more often and are more likely to diagnoses
  - Onset of disease may not be well documented
- Applications include:
  - Identifying unknown drug side effects
  - Repurposing existing drugs
  - Identifying treatment-resistant subgroups

Matched Case-Controls Design

- Case-control studies are a staple of observational data analysis
  - Cases = patients with a disease
  - Controls = patients without a disease
  - Exposure window = time preceding onset of a disease

- Simultaneously look at multiple risk factors
- Useful to initially establish an association between a risk factor and a disease or outcome

Analysis Pipeline



Experimental Setting

- Population: 69,121 patients consented in the Partners Healthcare Biobank
- Cases
  - Patients with a diagnosis of Osteoporosis (ICD-10: M80.*; M81.*; ICD-9 733.0*)
- Controls
  - No lifetime history of Osteoporosis
  - Matched using CEM on index_year, age_at_index, gender and race
- Exposures
  - All RxNorm drug ingredients/combinations) prescribed to at least 100 patients (901 drugs)
- Effect Estimates
  - Unadjusted Risk Ratio (riskratio.boot from R epitools package)
  - logit: Logistic Regression (adjusted for index_year and number of visits in window (log-adjusted)
  - clogit: Conditional logistic regression (adjusted for index_year, number of visits and matching stratum from R survival package)

Time Parameters

- Index Date:
  - Cases: First record diagnosis of Osteoporosis
  - Controls: Random visit date selected on the same year of the matched case
- Exposure Window:
  - 730 days (2 years) to 30 days prior to index date
  - Patients with no visits in the exposure window are excluded from both cases and controls

Effect Estimates

	Unexposed	Exposed
Osteoporosis	(**a**) NO Osteoporosis + Drug NOT prescribed in exposure window	(**b**) Osteoporosis + Drug NOT prescribed in exposure window
NO Osteoporosis	(**c**) NO Osteoporosis + Drug prescribed in Exposure Window	(**d**) Osteoporosis + Drug prescribed in exposure window

Unadjusted Risk Ratio (RR)

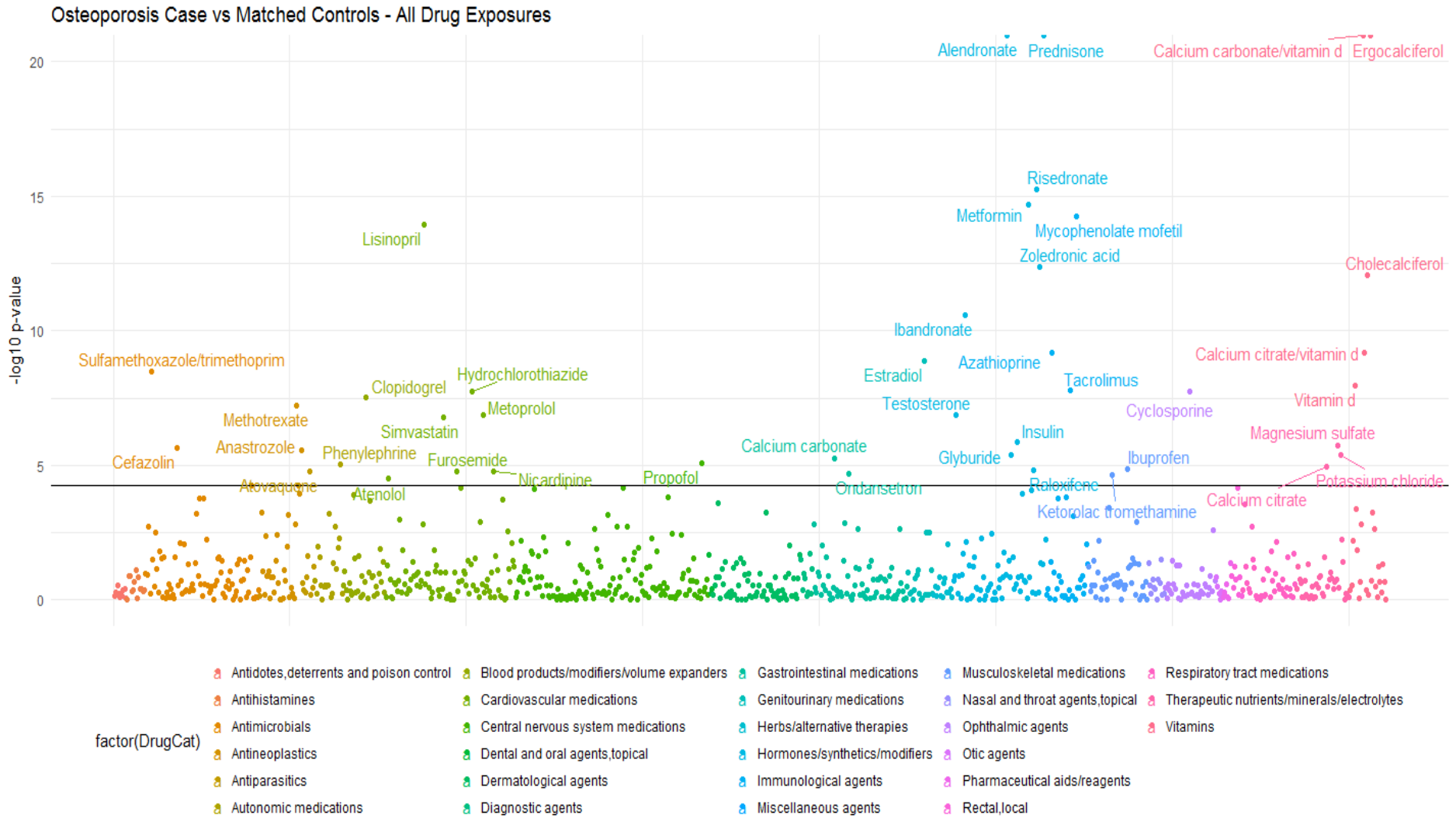
Logistic regression:

Drug Prescribed in Window (0/1) ~ Osteoporosis (0/1) + Year of Index Event + log(visits in exposure window)

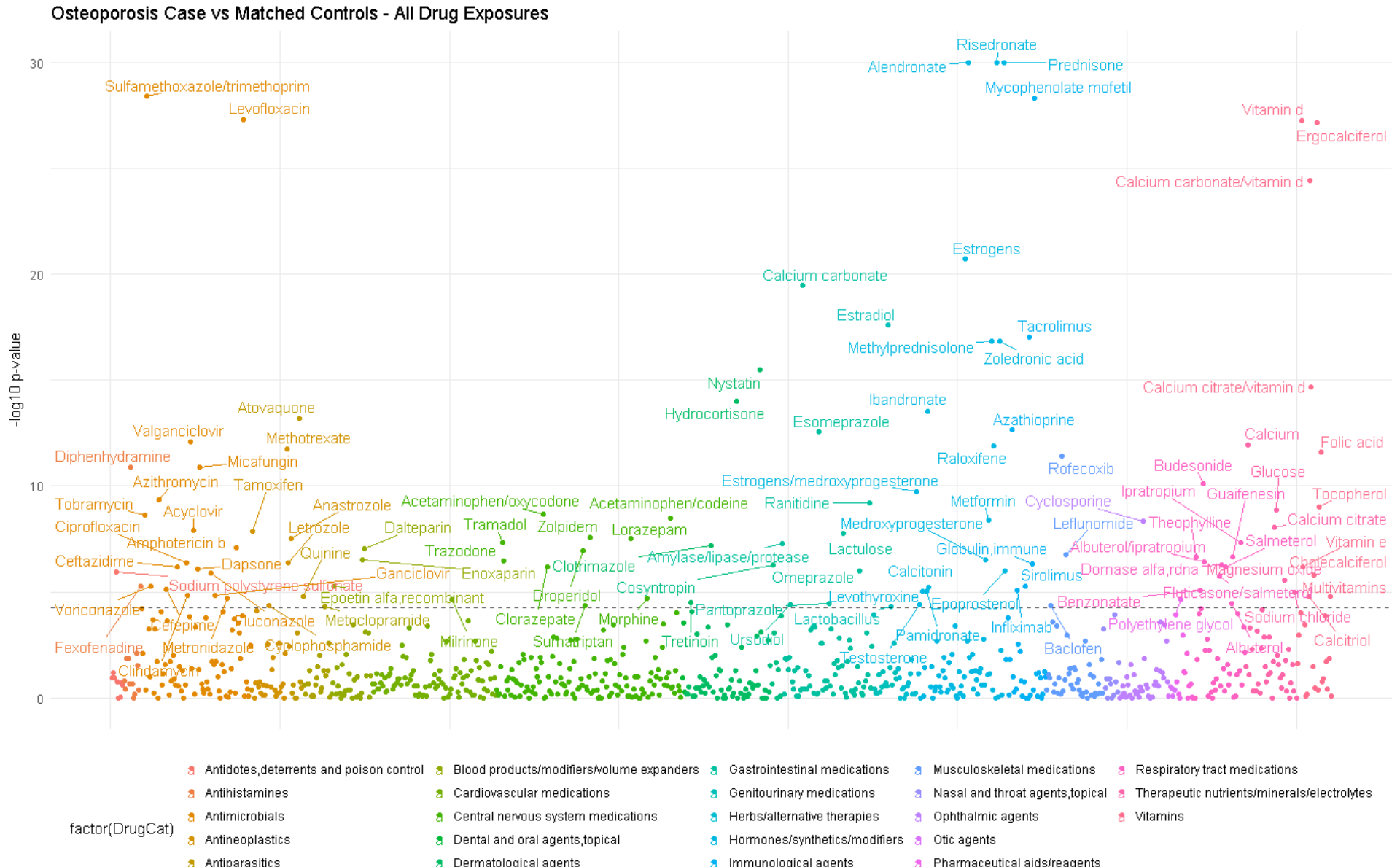
Conditional logistic regression:

Drug Prescribed in Window (0/1) ~ Osteoporosis (0/1) + Year of Index Event + log(visits in exposure window) + strata(match_strata)

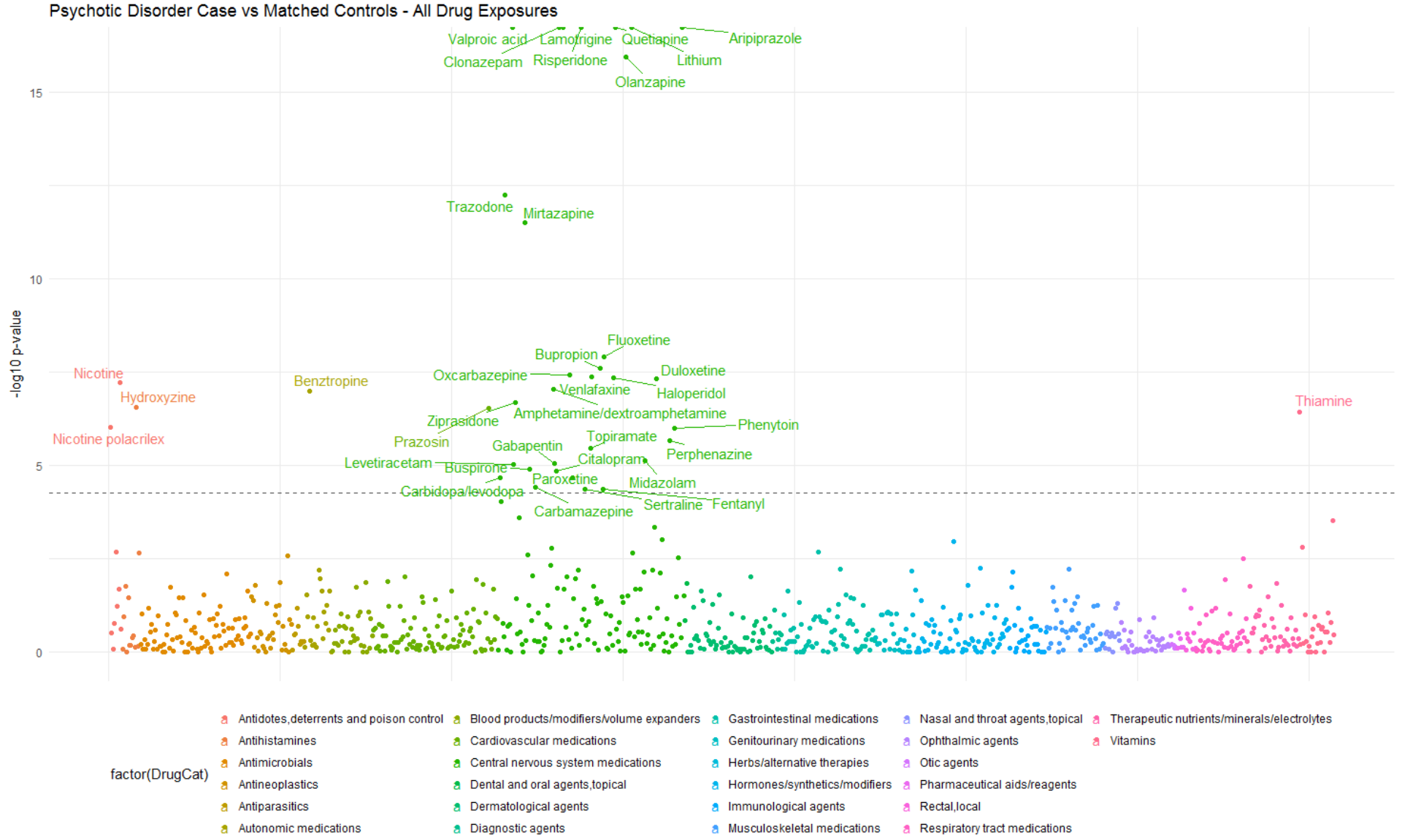
Results


Osteoporosis Case vs Matched Controls - All Drug Exposures
-log10 p-value
20
15
10
5
0
Alendronate
Prednisone
Calcium carbonate/vitamin d
Ergocalciferol
Risedronate
Metformin
Mycophenolate mofetil
Zoledronic acid
Cholecalciferol
Lisinopril
Ibandronate
Sulfamethoxazole/trimethoprim
Calcium citrate/vitamin d
Azathioprine
Estradiol
Tacrolimus
Hydrochlorothiazide
Clopidogrel
Metoprolol
Testosterone
Vitamin d
Cyclosporine
Methotrexate
Simvastatin
Insulin
Magnesium sulfate
Anastrozole
Phenylephrine
Calcium carbonate
Glyburide
Ibuprofen
Cefazolin
Furosemide
Nicardipine
Propofol
Potassium chloride
Atovaquone
Atenolol
Ondansetron
Raloxifene
Calcium citrate
Ketorolac tromethamine
factor(DrugCat)
Antidotes,deterrents and poison control
Antihistamines
Antimicrobials
Antineoplastics
Antiparasitics
Autonomic medications
Blood products/modifiers/volume expanders
Cardiovascular medications
Central nervous system medications
Dental and oral agents,topical
Dermatological agents
Diagnostic agents
Gastrointestinal medications
Genitourinary medications
Herbs/alternative therapies
Hormones/synthetics/modifiers
Immunological agents
Miscellaneous agents
Musculoskeletal medications
Nasal and throat agents,topical
Ophthalmic agents
Otic agents
Pharmaceutical aids/reagents
Rectal,local
Respiratory tract medications
Therapeutic nutrients/minerals/electrolytes
Vitamins

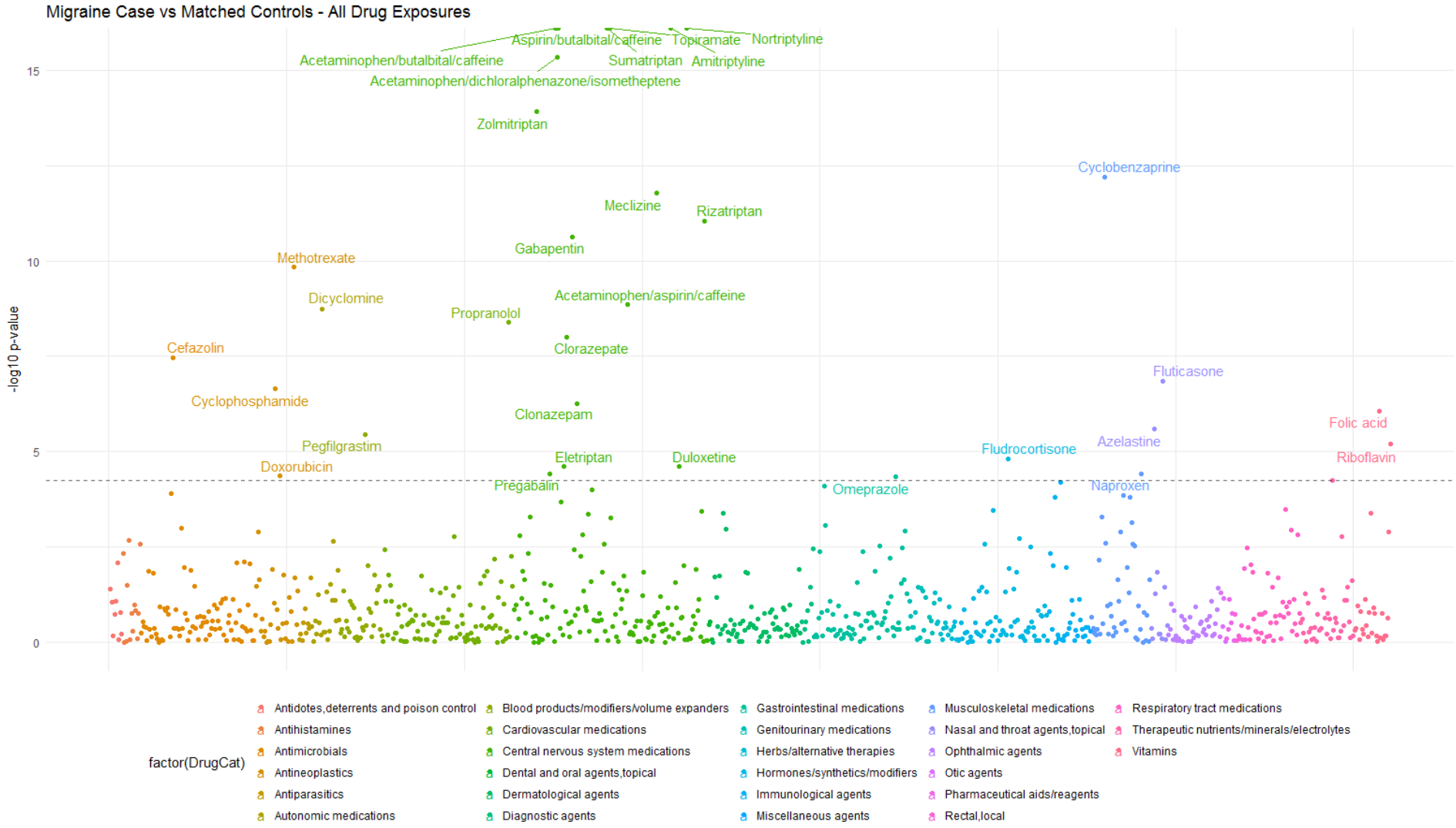
Results – No control for utilization



Results


Psychotic Disorder Case vs Matched Controls - All Drug Exposures
-log10 p-value
15
10
5
0
Valproic acid
Lamotrigine
Quetiapine
Aripiprazole
Clonazepam
Risperidone
Lithium
Olanzapine
Trazodone
Mirtazapine
Fluoxetine
Bupropion
Nicotine
Benztropine
Oxcarbazepine
Duloxetine
Venlafaxine
Haloperidol
Hydroxyzine
Thiamine
Amphetamine/dextroamphetamine
Ziprasidone
Phenytoin
Nicotine polacrilex
Topiramate
Prazosin
Gabapentin
Perphenazine
Levetiracetam
Citalopram
Buspirone
Paroxetine
Midazolam
Carbidopa/levodopa
Sertraline
Fentanyl
Carbamazepine
factor(DrugCat)
Antidotes,deterrents and poison control
Antihistamines
Antimicrobials
Antineoplastics
Antiparasitics
Autonomic medications
Blood products/modifiers/volume expanders
Cardiovascular medications
Central nervous system medications
Dental and oral agents,topical
Dermatological agents
Diagnostic agents
Gastrointestinal medications
Genitourinary medications
Herbs/alternative therapies
Hormones/synthetics/modifiers
Immunological agents
Musculoskeletal medications
Nasal and throat agents,topical
Ophthalmic agents
Otic agents
Pharmaceutical aids/reagents
Rectal,local
Respiratory tract medications
Therapeutic nutrients/minerals/electrolytes
Vitamins

Results


Migraine Case vs Matched Controls - All Drug Exposures
-log10 p-value
15
10
5
0
Aspirin/butalbital/caffeine
Topiramate
Nortriptyline
Acetaminophen/butalbital/caffeine
Sumatriptan
Amitriptyline
Acetaminophen/dichloralphenazone/isometheptene
Zolmitriptan
Cyclobenzaprine
Meclizine
Rizatriptan
Gabapentin
Methotrexate
Dicyclomine
Acetaminophen/aspirin/caffeine
Propranolol
Cefazolin
Clorazepate
Fluticasone
Cyclophosphamide
Clonazepam
Folic acid
Pegfilgrastim
Azelastine
Fludrocortisone
Eletriptan
Duloxetine
Riboflavin
Doxorubicin
Pregabalin
Omeprazole
Naproxen
factor(DrugCat)
Antidotes,deterrents and poison control
Antihistamines
Antimicrobials
Antineoplastics
Antiparasitics
Autonomic medications
Blood products/modifiers/volume expanders
Cardiovascular medications
Central nervous system medications
Dental and oral agents,topical
Dermatological agents
Diagnostic agents
Gastrointestinal medications
Genitourinary medications
Herbs/alternative therapies
Hormones/synthetics/modifiers
Immunological agents
Miscellaneous agents
Musculoskeletal medications
Nasal and throat agents,topical
Ophthalmic agents
Otic agents
Pharmaceutical aids/reagents
Rectal,local
Respiratory tract medications
Therapeutic nutrients/minerals/electrolytes
Vitamins

Conclusion

- Pay attention to the facts
- A case-control approach applied to large EHR datasets can identify true causal effects of drug exposures with the aim of monitoring the safety of medications and identifying candidates for drug repurposing.
- The Ri2b2casecontrol tools implements the methods in an i2b2 framework.
- Future efforts will be aimed at minimizing bias due to missing data and inaccurate outcome definitions
- Longer-term goal of incorporating genomic and environmental data into methods
- Improve data workflow with UI within the i2b2 webclient

Ri2b2casecontrol R package

https://github.com/vcastro/Ri2b2casecontrol

case_control {Ri2b2casecontrol} R Documentation

case_control

Description

The main case control function to run a case control analysis from i2b2-generated data files

Usage

```
case_control(p, d, dict, exposure_cds, visit_cd = "V", outcome_cd = "O",
  riskWindow_daysPreIndex_start = 365, riskWindow_daysPreIndex_end = 1,
  controls_num = 3)
```

Arguments

p	A data frame of patients demographics
d	A data frame of patient data in a tall file
dict	A data frame of descriptions for each concept_cd in the data_tbl
exposure_cds	A string vector of exposures codes to test
visit_cd	The code used for visits (default is V)
outcome_cd	The code used for the outcome in the data_tbl (default is O)
riskWindow_daysPreIndex_start	The number of days prior to the index date to start the risk window (default is 365)
riskWindow_daysPreIndex_end	The number of days prior to the index date to end the risk window (default is 1)
controls_num	Number of controls to match to each case

Value

Resources

- Standalone R package
  - Ri2b2casecontrol
    https://github.com/vcastro/Ri2b2casecontrol
  - Ri2b2matchcontrols (implements CEM based on R cem package)
    https://github.com/vcastro/Ri2b2matchcontrols
- https://rc.partners.org/
- https://i2b2.org/
- Questions:
  vcastro@partners.org