Session III: The Landscape of Signal Detection Approaches for Longitudinal Data

Join the conversation with #sentinelinitiative

Landscape Overview of Signal Detection Techniques

Martin Kulldorff
Division of Pharmacoepidemiology and Pharmacoeconomics
Brigham and Women's Hospital and Harvard Medical School


WHEN WE TESTED THIS DRUG ON MICE, NOBODY NOTICED ANY SIDE EFFECTS.
SCHWADRON

Data Mining

Simultaneous evaluation of multiple drug / adverse event pairs.

Common goals of data mining methods

- Find unknown adverse reactions, if they exist
- Few false positives, or else, easily explained false positives
- Sufficient power to detect rare adverse reactions
- (Find known adverse reactions, if any)

Data Mining Design Features

- Data mining methods have different features.
- These can be combined freely to create hybrid approaches.
- So, approach it like ordering ice cream, picking your favorite cone, ice cream flavors and toppings.

Key Feature: Data Source

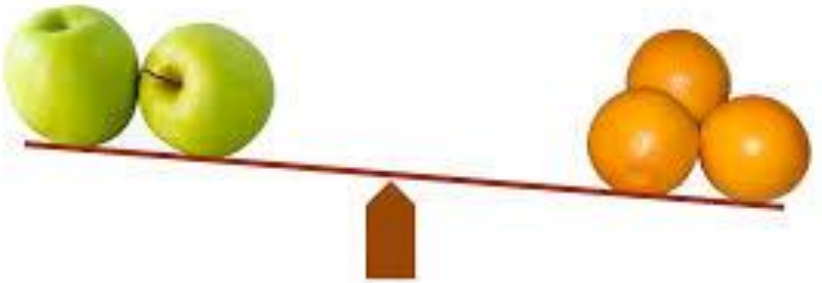
- Spontaneous reports (e.g. AERS)
- Clinical trials
- Disease registries
- Electronic health records
- Insurance claims data
- …

Key Feature: Risk Window

- Risk window identical to exposure period
- One or more pre-specified risk window, such as 1-14 days after initial exposure
- Temporal scan, simultaneously evaluating hundreds of potential risk windows

Key Feature: Comparison Group

- All individuals
- Healthy individuals
- Users of all other drugs
- Users of a similar drug



- Self-controls, pre-exposure control window
- Self-controls, post-exposure control windows

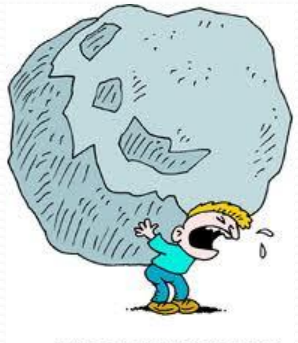
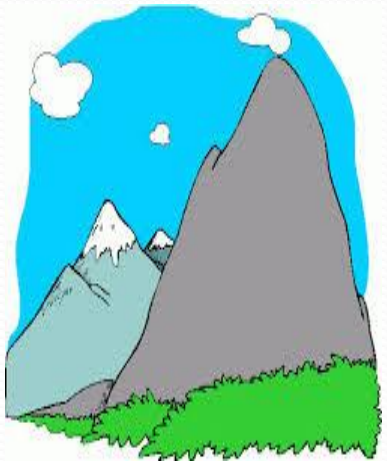
Key Feature: Covariate Adjustment

- None
- Age, gender, calendar time, geography, etc
- Concomitant exposures
- Propensity score matching

Key Feature: Outcome Granularity

- Use collection of very specific diagnoses, such as ICD-9 codes (e.g. acute liver failure).
- Use smaller collection of more general groups of related diagnoses (e.g. liver disease)
- Simultaneously use both of the above, plus intermediate levels



Key Feature: Multiple Testing Adjustment

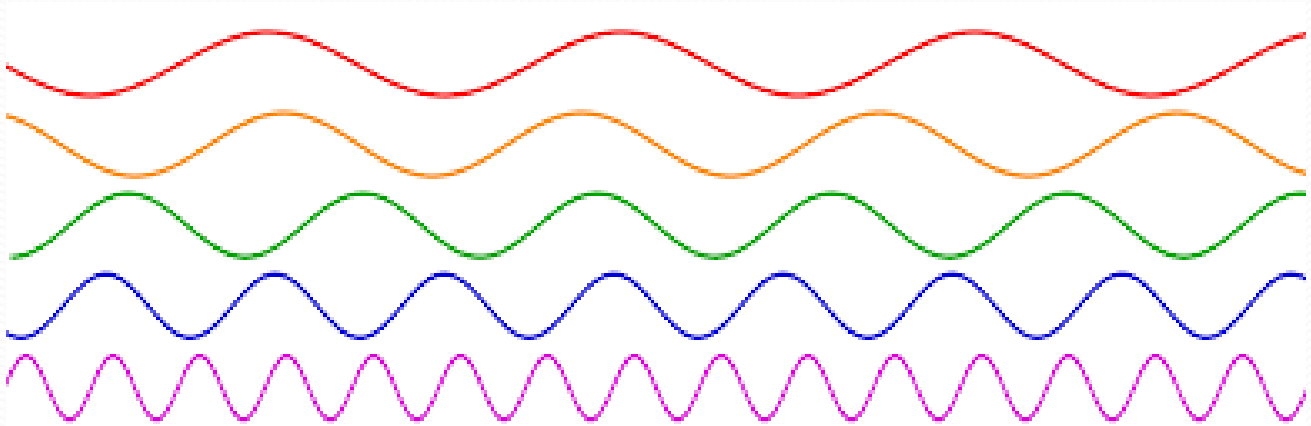
- No adjustment
- Informal adjustment, such as lower 95% CI >2
- Formal analytical Bonferroni type adjustment
- Formal Monte Carlo adjustment, with random data generated under the null hypothesis

Key Feature: Effect Estimates

- Relative Risk / Odds Ratio
- Empirical Bayes Shrinkage Estimates
- Attributable Risk / Risk Difference

Key Feature: Frequency

- One single look at the data
- Multiple looks over time, as more data accrues
- Real-time safety surveillance



Key Feature: Size and Type of Net

- One drug, thousands of disease outcomes
- One disease, thousands of drugs
- Thousands of drugs and thousands of outcomes
- Specific population, such as pregnant women and birth defects
- Drug-drug interactions

Signal Detection using TreeScan

Judith C. Maro
Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA

Tree-Based Scan Statistics are Enabled by:

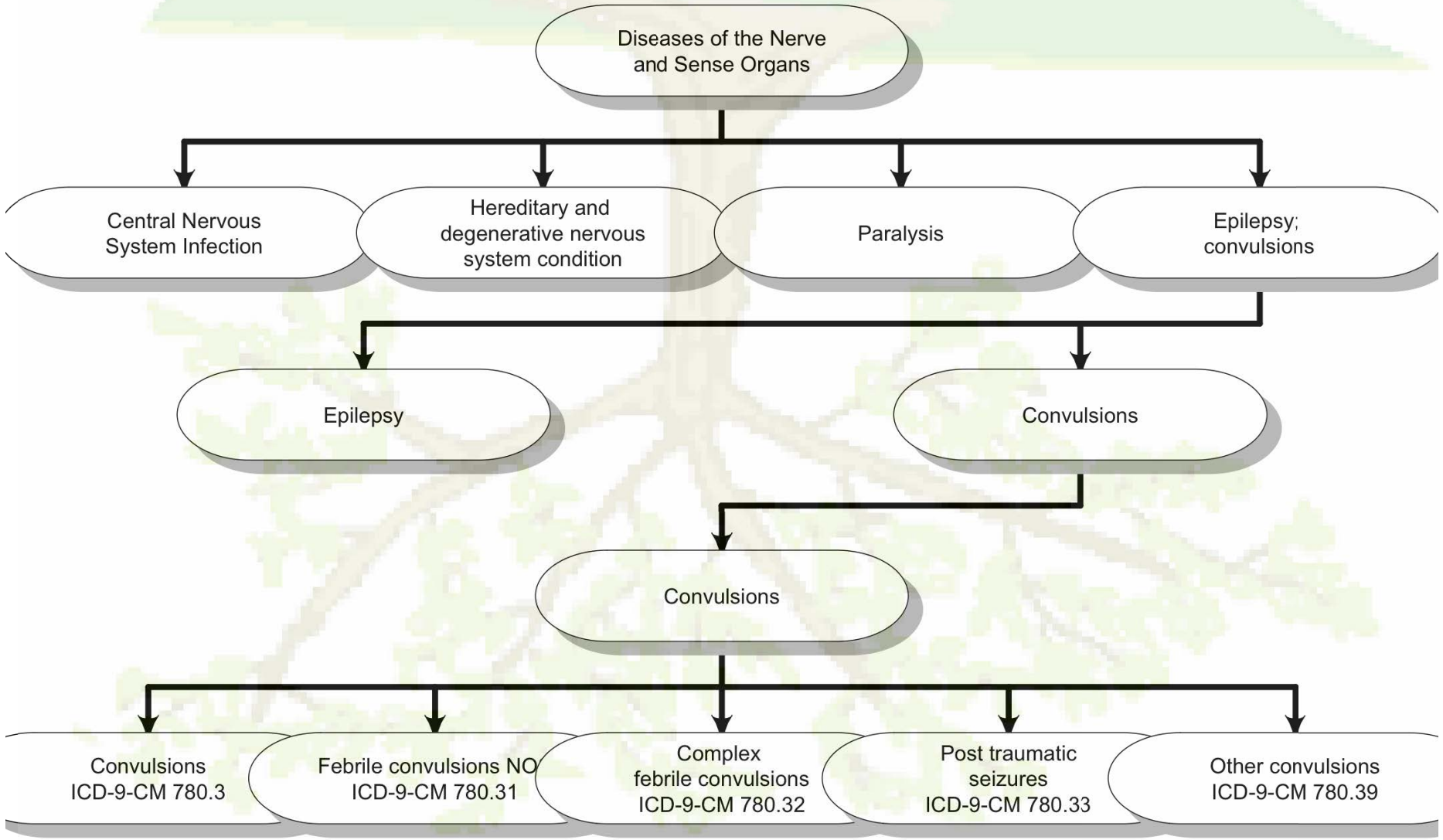
- A signal detection / data-mining method
- Scans electronic health outcome data that are grouped into hierarchical tree structures
- Automatically adjusts for multiple hypothesis testing

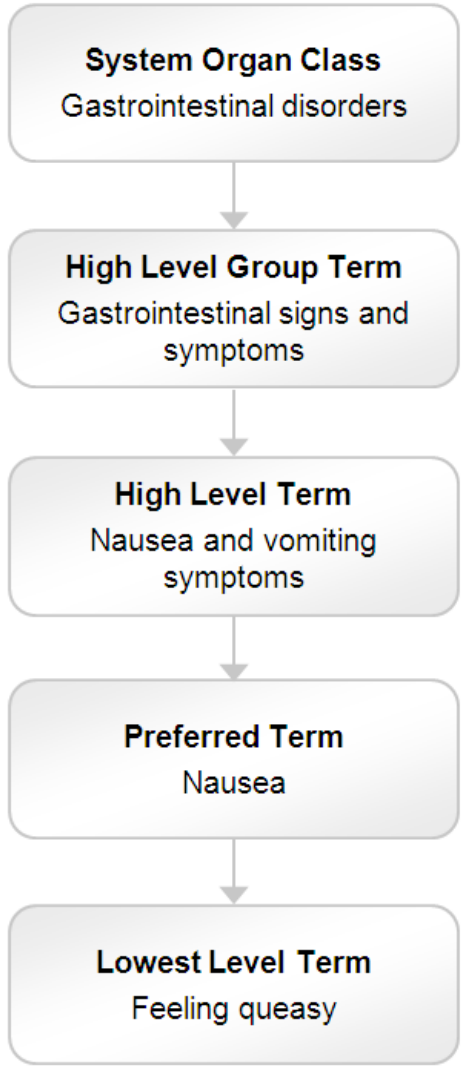
http://www.treescan.org

Data-Mining Designs with Trees

- Exposure-Oriented - 1 Exposure: N Outcomes
  - Uses Multi-Level Clinical Classification System (MLCCS) where N=~8000
- Outcome-Oriented - M Exposures: 1 Outcome
  - Uses Medi-Span Therapeutic Classification System (Drug Tree) where M=300,000+
- *Future - M Exposures: N Outcomes*

Data Arranged in a Tree Structure





Study Designs Compatible with TreeScan Analytics

Sentinel

		TreeScan Analytics					
		Poisson Model		Bernoulli Model		Tree-Temporal Model	
		Unconditional	Conditional	Unconditional	Conditional	Unconditional	Conditional
Study Designs	Self-Controlled Design			X	X	X	X
	Propensity Score or other Fixed Ratio Match Design			X			
	Stratified Cohort Design	X	X				

Unconditional means the null hypothesis relies on an external input about the expected outcomes.
Conditional means the null hypothesis is determined by the characteristics of the incoming data set.

How has TreeScan been evaluated thus far?

Simulated Datasets

- **Advantages**
  - Artificially inject "excess risk" of variable specific sizes
  - Allows quantitative assessment of method under "experimental conditions" where "truth is known"
- **Limitations**
  - Simulated data has a range of realistic representations. Early simulations are quite artificial

Empiric Assessments

- **Advantages**
  - Empiric testing with real data
  - Allows assessment of method under real life conditions
  - Can be effective method to assess performance if test case is well characterized
- **Limitations**
  - Can be challenging to interpret unexpected results
  - Need additional information to investigate unexpected results

Self-Controlled Designs (Tree-Temporal)

Sentinel

Sentinel

PDS Pharmacoepidemiology & Drug Safety

Official Journal of the International Society for Pharmacoepidemiology

ISPE

ORIGINAL RE

Mening
Vaccine
method

Rongxia Li
Stanley Xu,

First publish

Assessment of Quadrivalent Human Papillomavirus Vaccine Safety Using the Self-Controlled Tree-Temporal Scan Signal-Detection Method in the System

W Katherine Yih, Judith C Maro, Michael Nguyen, M
Carolyn Balsbaugh, David V Cole, Inna Dashevsky, Ada
Martin Kulldorff

American Journal of Epidemiology, Volume 187, Issue
1269–1276, https://doi.org/10.1093/aje/kwy023
Published: 23 February 2018 **Article history**

758 | Signal detection using tree-temporal scan statistics in the sentinel system

Judith C. Maro[1]; Michael D. Nguyen[2]; Rima Izem[2]; Ya-Hui Hsueh[2]; Inna Dashevsky[1]; Austin Cosgrove[1]; S. Christopher Jones[2]; Jacqueline M. Major[2]; Esther H. Zhou[2]; Elande Baro[2]; Joshua J. Gagne[3]; Shirley V. Wang[3]; Martin Kulldorff[3]

[1] *Harvard Medical School and Harvard Pilgrim Healthcare Institute, Boston, Massachusetts;* [2] *Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland;* [3] *Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts*

Propensity Score Matched Designs

Epidemiology. 29(6):895–903, NOV 2018
DOI: 10.1097/EDE.0000000000000907, PMID: 30074538
Issn Print: 1044-3983
Publication Date: 2018/11/01

Print

Data Mi
Propens
Statistic

Shirley V. Wang; Ju
Gagne; Elisabetta
Sebastian Schnee

+ Author Informat

Submit Comment

Development and Evaluation of a Global Propensity Score for Data Mining with Tree-Based Scan Statistics

Project Title	**Development and Evaluation of a Global Propensity Score for Data Mining with Tree-Based Scan Statistics**
Date Posted	*Friday, August 10, 2018*
Status	In progress
Deliverables	Development and Evaluation of a Global Propensity Score for Data Mining with Tree-Based Scan Statistics: Protocol

Stratified Cohort Designs with Referent Cohort

PDS Pharmacoepidemiology & Drug Safety

Official Journal of the International Society for Pharmacoepidemiology

ISPE

Original Repor

Drug safe

Martin Kulldor
Richard Platt,

First published

Read the full t

Pharmaceutics **2013**, *5*(1), 179-200; https://doi.org/10.3390/pharmaceutics5010179

Open Access Article

Drug Advers
the Gamma
Tree-based S

Jeffrey S. Brown [1,2,*]
Inna Dashevsky [1],
K. Arnold Chan [5,6],
Lisa Herrinton [2,10],
David Smith [2,13] an

eGEMs AcademyHealth The Journal for Electronic Health Data and Methods

Start Submi

Reading: Statistical Power for Postlicensure Medical Product Safety Data Mining

Share:

Empirical research

Statistical Power for Postlicensure Medical Product Safety Data Mining

Authors: Judith C. Maro, Michael D. Nguyen, Inna Dashevsky, Meghan A. Baker, Martin Kulldorff

Strengths of TreeScan

1. Takes advantage of hierarchical nature of clinical concepts in the form of a tree structure.
2. Investigator does not need to understand how particular outcomes are coded (i.e., can be indifferent to the granularity of the outcome data)
3. Formal control for multiple hypothesis testing (Overall Type 1 error)

Limitations of TreeScan

1. All outcomes are treated identically across the tree (8000+) regardless of their time of onset, severity, etc.
2. Complex outcomes (algorithms such as 2 codes within X days of each other) are not tested with TreeScan.
3. Individual study designs have limitations depending on the design chosen.

Signal Detection in the Sentinel System

Washington, DC
Dec 3, 2018

ORACLE
Health Sciences

Bayesian Shrinkage and Meta-Analysis: Possible Applications to Sentinel

William DuMouchel, PhD
Chief Statistical Scientist
Oracle Health Sciences
Miami, FL, USA

ORACLE

Bayesian Shrinkage Techniques and Meta-Analysis

Description and Scenario

These techniques reduce variation by combining multiple examples to allow borrowing strength after estimating a prior distribution of average effects

They are most effective when the study design seeks estimates for many parallel problems

A Bayesian model assumes similarity of effects or sources of variation across the multiple problems

Data Requirements, Strengths and Limitations

These methods work best with large databases to be able to draw on multiple estimations and measure variance components

A primary strength is that they often provide accurate adjustments for multiple comparisons conundrums that are especially vexing for safety analyses
[Accomplished by estimation of assumed prior effect variances across examples]

Results can mislead if individual problems are black swan outliers that don't follow the general pattern of most other examples.

Shrinking Safety Signals Toward Class Effects

- Observational Database for Drug Adverse Reactions
  - N = Counts of Drug-Event Combinations (DECs)
  - E = Expected Counts Based on some No-Effect Model
- Null-Hypothesis Models for Expected Counts
  - Adjustment for Age and Gender and other Covariates
  - Adjustment for Concomitant Drugs (Large Scale Regression)
  - Longitudinal exposure models
- Two-way Shrinkage Model
  - Assume analysis of prespecified set of DECs
  - A class of drugs for the same indication
  - A set of medically similar adverse effects

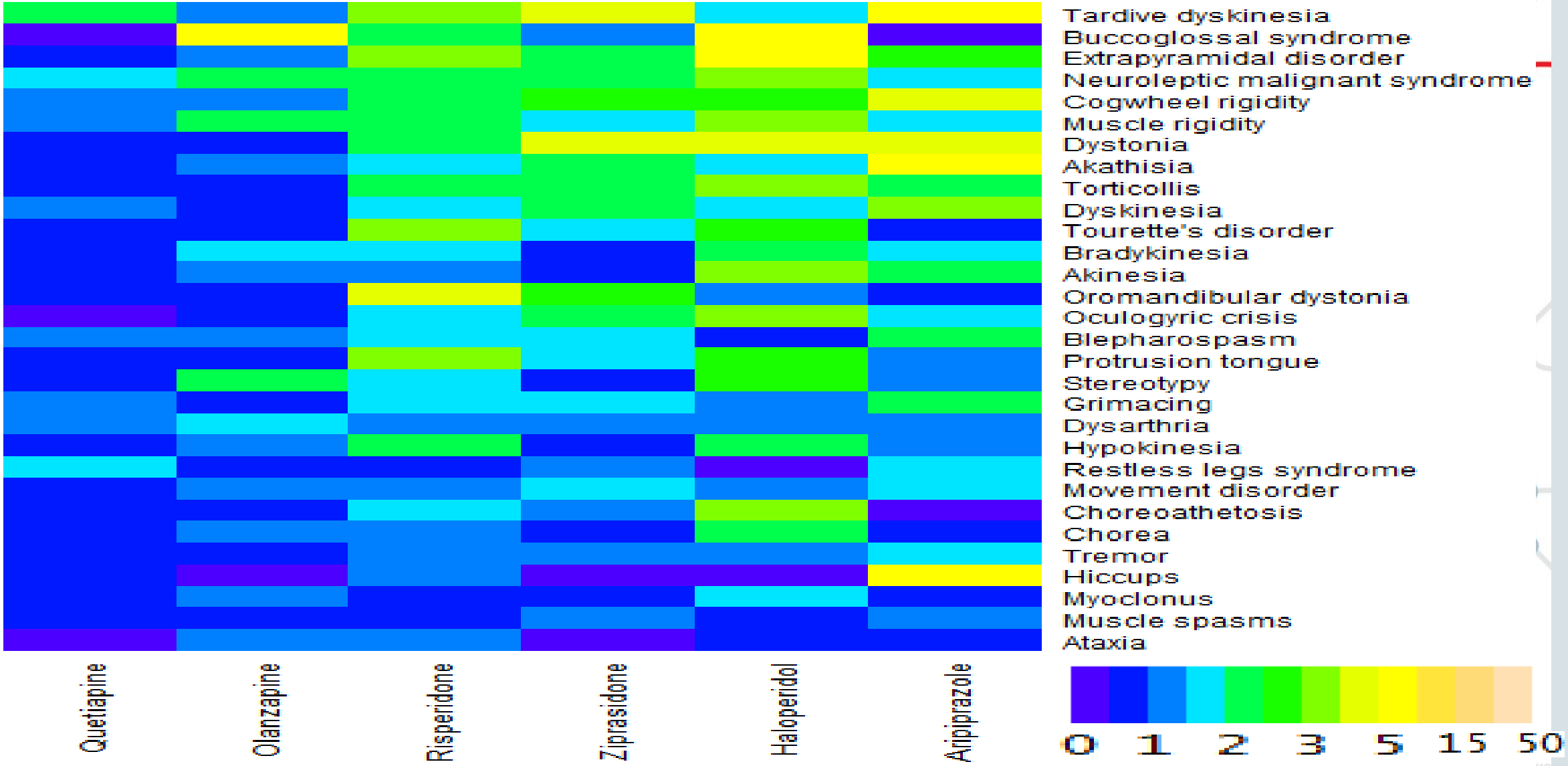
Rationale for Two-Way Shrinkage

- Similar Drugs may have a Class Effect on each PT
- Similar Mechanisms Affect Medically Related Events
  - PTs Close in the MedDRA Hierarchy May Have Common Causes
- There Are Probably Specific Drug-Event Associations
  - But We Have Noisy Measurements for Rare Combinations
  - Estimate Deviations from Overall Drug and Event Patterns
- Decompose Associations: Drug Effect × PT Effect × Residual (Interaction) Effect
  - Prior Distributions Can Shrink All 3 Types of Effects Toward 1
  - High-Variance Estimates Will Get Shrunk the Most
- DuMouchel W, Harris JE (1983) *JASA* **78**:293-315
  - Fits Similar Model to a Collection of Environmental Dose-Response Slope Estimates

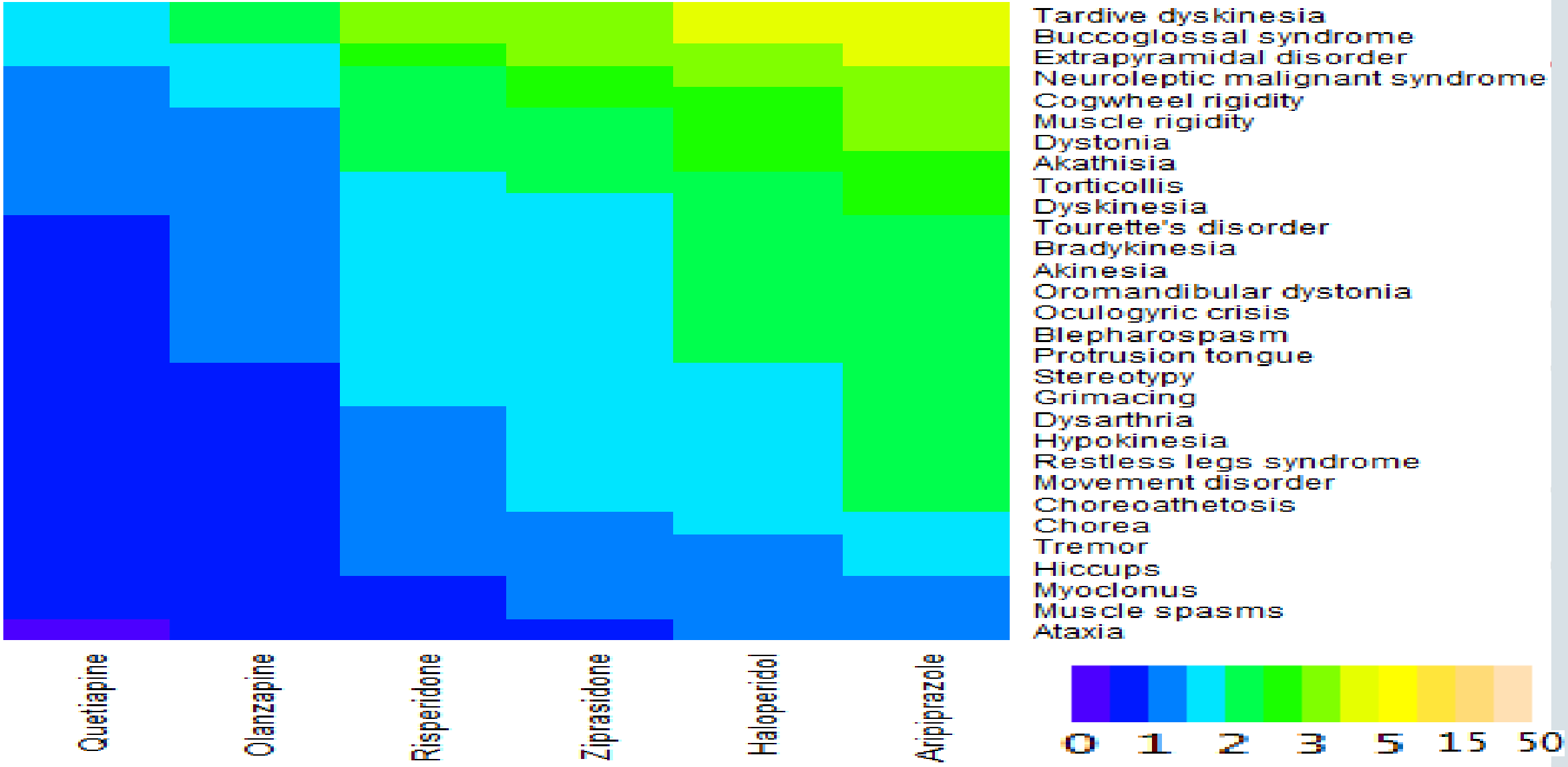
Antipsychotics and Movement Disorders in FAERS

- Drugs—J = 6 Antipsychotics
  - Aripiprazole Haloperidol Olanzapine Quetiapine Risperidone Ziprasidone
- Events—K = 30 PTs for Movement Disorders
  - Selected as the 30 PTs Related to Movement Disorder that Had the Greatest Total No. of Reports across all 6 Drugs
- Counts N_{jk} and Expected Counts E_{jk} [Drug j, Event k]
- Poisson Regression Estimates Average Drug and PT Trends
  - $N_{jk} \sim \text{Poisson}(E_{jk} \exp\{\alpha_j + \beta_k\})$ [Assume one $\beta_k = 0$ to normalize]
- Shrink Observed N_{jk}/E_{jk} Toward the Overall Regression Trends

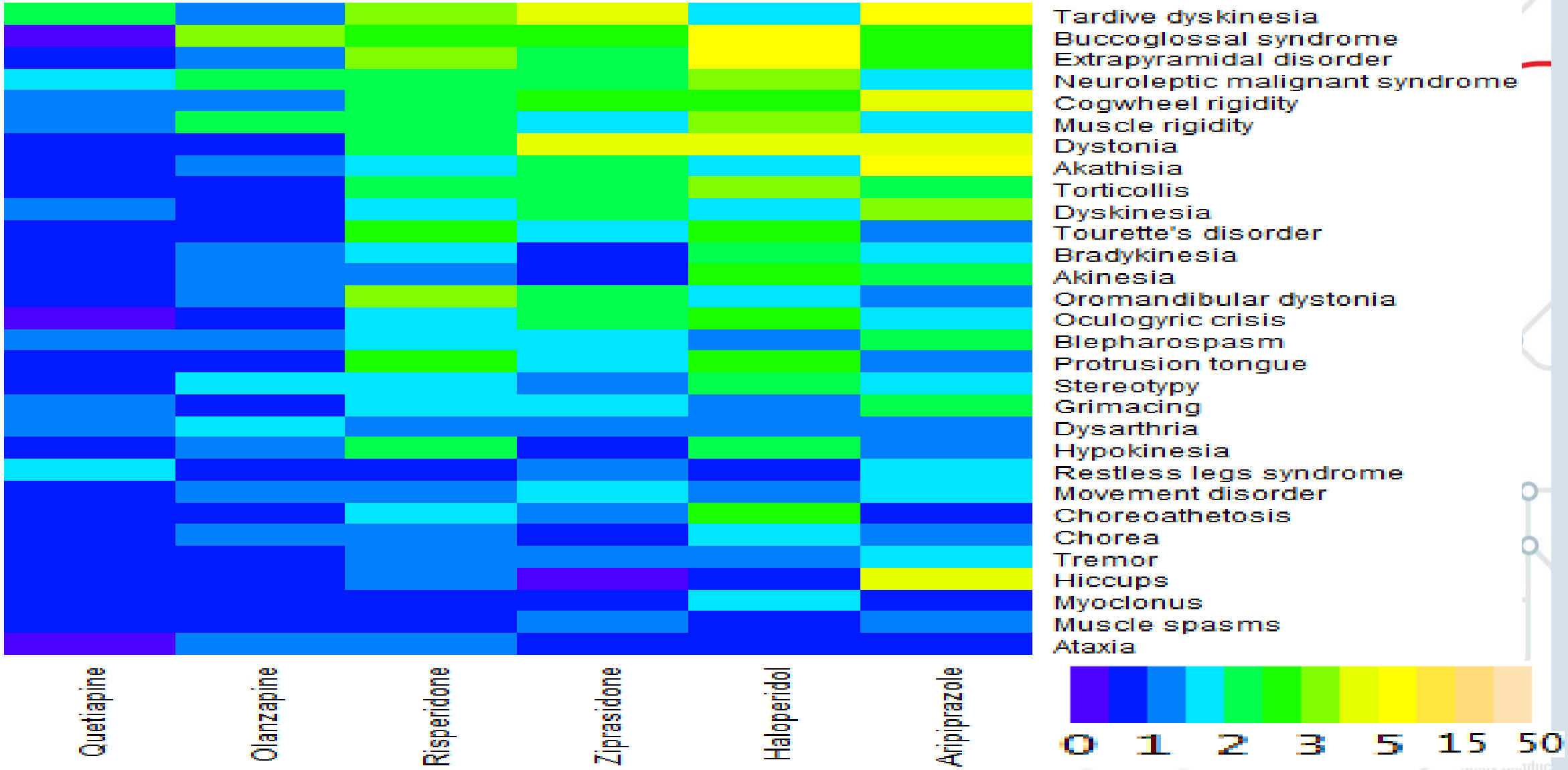
Ratios N/E [E from RGPS]



Regression Fit to N/E



Two-Way Shrinkage of N/E



Shrinkage Model Across Different Data Sources

- Let N_{djk} and E_{djk} Denote Observed and Expected Counts
  - Database d, Drug j and Event k
- $\lambda_{djk} > 0$
  - "True" disproportionality of drug j and event k in database d
  - $N_{djk} \sim \text{Poisson}(E_{djk}\,\lambda_{djk})$
- Assume $\lambda_{djk} = \alpha_{dj} \times \beta_{dk} \times \gamma_{jk}$, where
  - $\alpha_{dj} \sim \text{Gamma}(A, A)$ so that each α_{dj} has prior mean equal to 1 and prior variance 1/A
  - $\beta_{dk} \sim \text{Gamma}(B, B)$
  - $\gamma_{jk} \sim \text{Gamma}(C, C)$
- Estimates of A, B, C and λ_{djk} Provide a Meta-Analysis of the Safety Situation
  - Several other models for λ_{djk} may be appropriate

Large-Scale Evidence Generation and Evaluation in a Network of Databases (LEGEND)

Martijn Schuemie

Janssen Research and Development – Global Epidemiology

UCLA - Biostatistics

Observational Data

Can be used to estimate

- Counts: how often does outcome occur in exposed?
- Associations: is exposure associated with greater counts?
- Causal effects: does an exposure increase the risk of an outcome?

Our goal: Signal = Causal effect

"Hypothesis free"

actually means

"Many hypotheses"

For example:

- a new drug – 'all' outcomes
- a class of drugs – class of outcomes

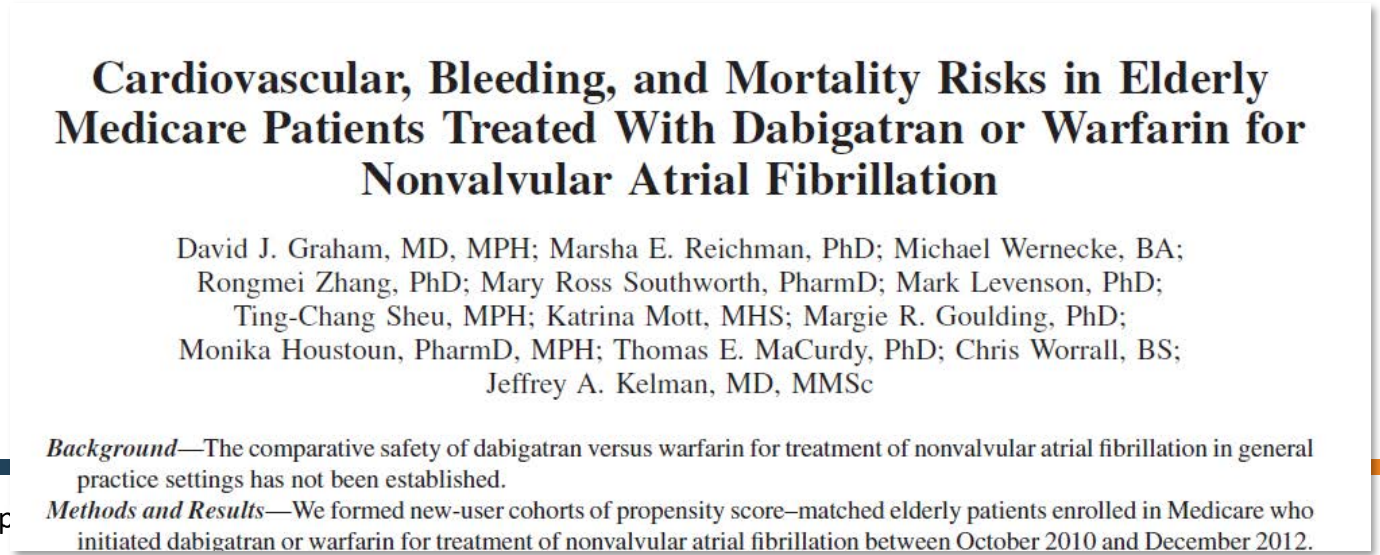
Estimating causal effects

Many methods exist, e.g.

- New-user cohort method using propensity score adjustment
- Self-Controlled Case Series (SCCS)
- Case-control
- Case-crossover
- Self-controlled cohort

Best practice for estimating causal effects in observational data

- Prespecify
- Transparency: protocol + source code
- Proper outcome definitions
- Sensitivity analyses
- Study diagnostics
- Multiple databases



Cardiovascular, Bleeding, and Mortality Risks in Elderly Medicare Patients Treated With Dabigatran or Warfarin for Nonvalvular Atrial Fibrillation

David J. Graham, MD, MPH; Marsha E. Reichman, PhD; Michael Wernecke, BA; Rongmei Zhang, PhD; Mary Ross Southworth, PharmD; Mark Levenson, PhD; Ting-Chang Sheu, MPH; Katrina Mott, MHS; Margie R. Goulding, PhD; Monika Houstoun, PharmD, MPH; Thomas E. MaCurdy, PhD; Chris Worrall, BS; Jeffrey A. Kelman, MD, MMSc

Background—The comparative safety of dabigatran versus warfarin for treatment of nonvalvular atrial fibrillation in general practice settings has not been established.

Methods and Results—We formed new-user cohorts of propensity score–matched elderly patients enrolled in Medicare who initiated dabigatran or warfarin for treatment of nonvalvular atrial fibrillation between October 2010 and December 2012.

LEGEND

LARGE-SCALE EVIDENCE GENERATION AND EVALUATION IN A NETWORK OF DATABASES

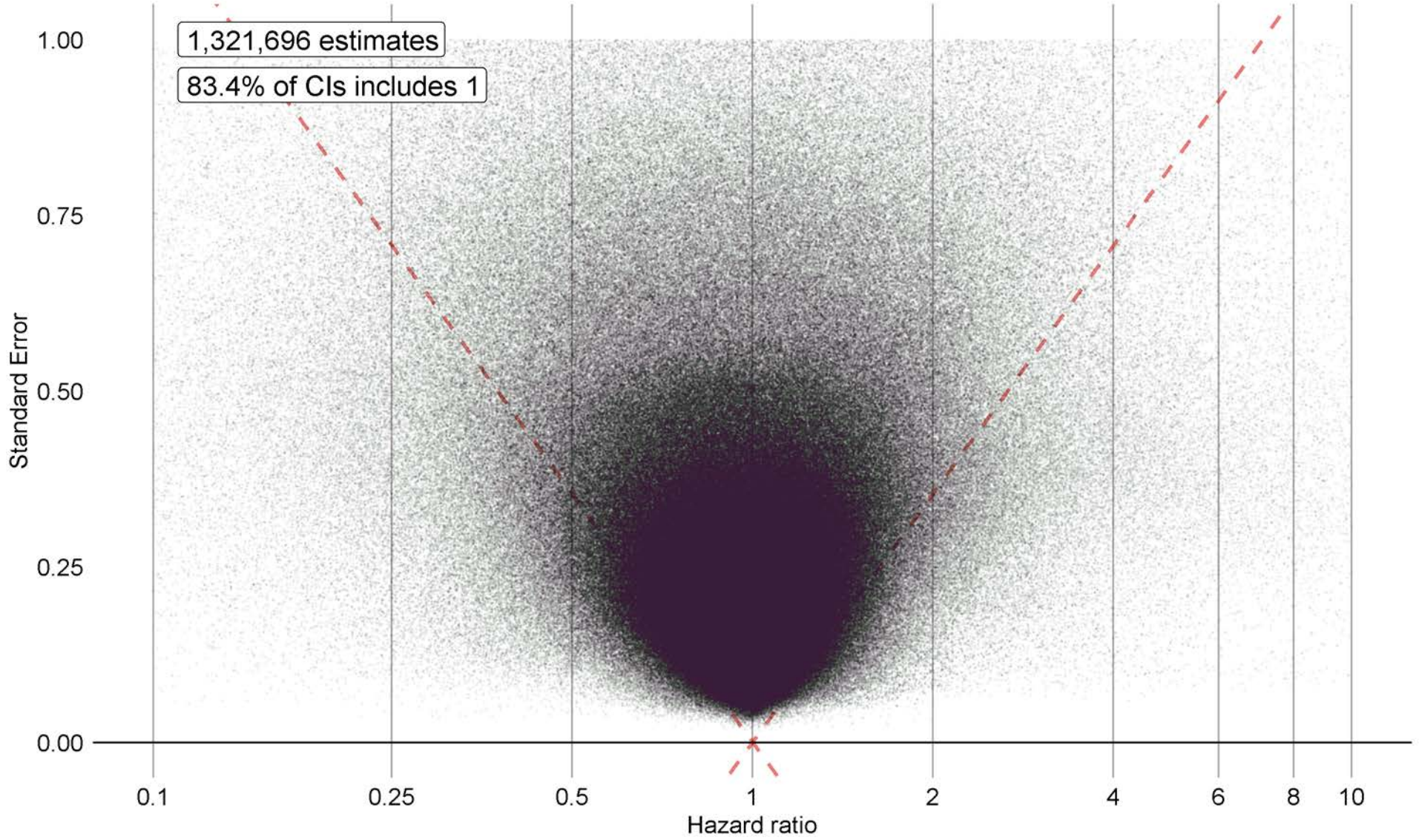
Hypertension treatments

- 10,278 comparisons between drugs, classes, and combinations of these
- 58 outcomes of interest
- 587,020 research questions

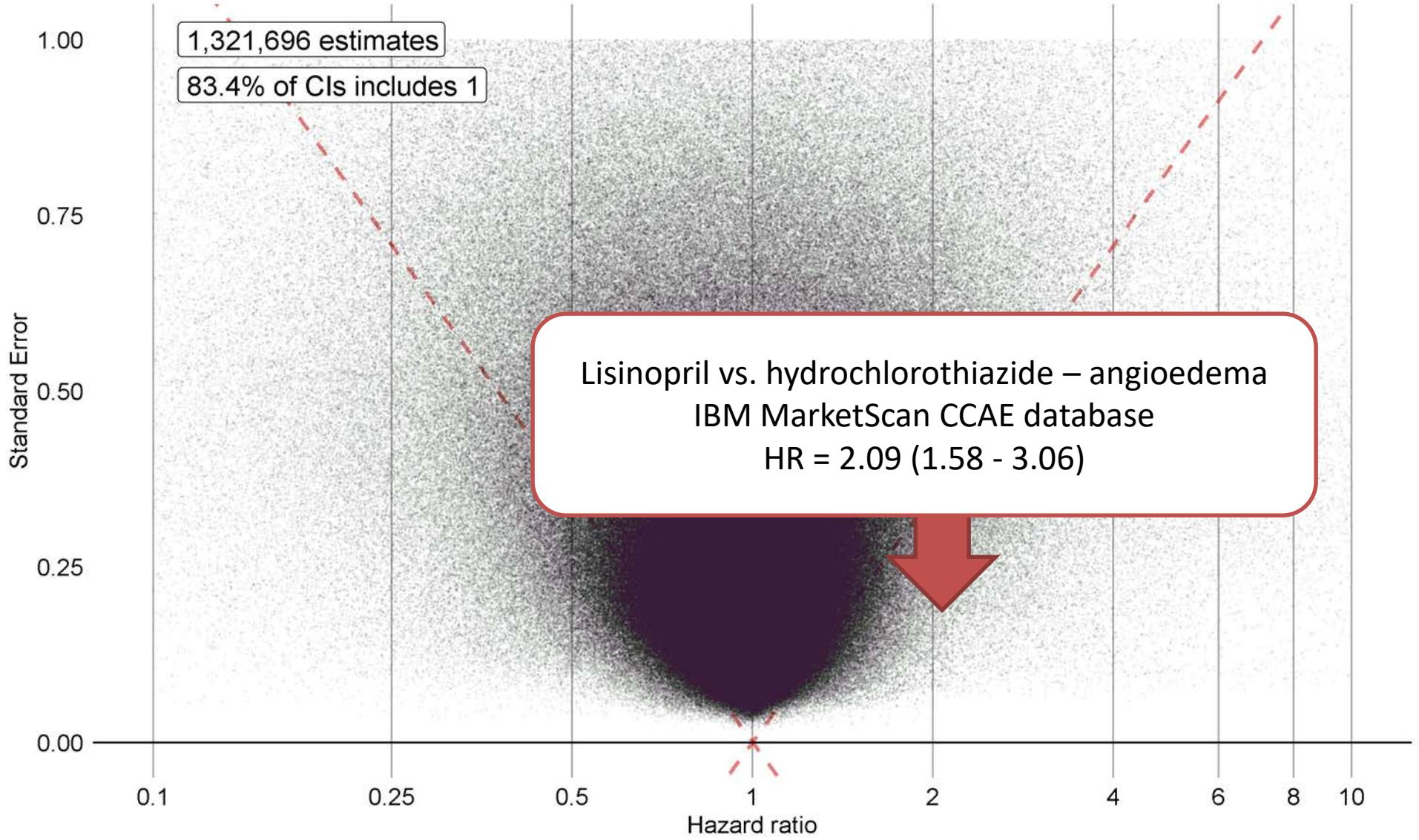
Methods

- New-user cohort design
- Large-scale propensity scores
- Proper outcome definitions
- 9 databases across the globe

LEGEND results



LEGEND results



Each analysis has all the content you should expect in a manuscript

Available for each of the 1,321,696 estimates

Angioedema risk in new-users of Lisinopril versus Hydrochlorothiazide for hypertension in the CCAE database

Martijn J. Schuemie[a,b,c], Patrick B. Ryan[a,b,d], Seng Chan Y
Marc A. Suchard[a,e,f,g]

[a]Observational Health Data Sciences and Informatics, New York, NY, USA; [b]Jan
California, Los Angeles, CA; [c]Department of Biomedical Informatics, Columbia
South Korea; [d]Janssen Institute, University of South Australia, Adelaide SA, Aus
Biomathematics, University of California, Los Angeles, CA; [f]Department of Hum

This report was **automatically** compiled on November 28, 2018.

**We conduct a large-scale study on the incidence of angioede among new users of lisinopril and hydrochlorothiazide from 2 to 2018 in the CCAE database. Outcomes of interest are estima of the hazard ratio (HR) for incident events between compara new users under on-treatment and intent-to-treat risk window sumptions. Secondary analyses entertain possible clinically relev subgroup interaction with the HR. We identify 647212 lisinopril 286973 hydrochlorothiazide patients for the on-treatment design, taling 465529 and 200649 patient-years of observation, and 671 160 events respectively. We control for measured confounding ing propensity score trimming and stratification or matching ba on an expansive propensity score model that includes all measu patient features before treatment initiation. We account for unm sured confounding using negative and positive controls to estim and adjust for residual systematic bias in the study design and d source, providing calibrated confidence intervals and p-values. terms of angioedema, lisinopril has a higher risk as compared to drochlorothiazide [HR: 2.09, 95% confidence interval (CI) 1.58 - 3.

new-user cohort design | comparative effectiveness | drug safety

The Large-scale Evidence Generation and Evaluation i Network of Databases (LEGEND) project aims to g erate reliable evidence on the effects of medical interventi using observational healthcare data to support clinical d sion making. LEGEND follows ten guiding principles (Supporting Information); chief among these stand that generate evidence at large-scale to achieve completeness a facilitate analysis of the overall distribution of effect size timates across treatments and outcomes. We also gener evidence consistently by applying a systematic approach acr all research questions and disseminate evidence regardless the estimates effects to avoid publication bias. These aims h overcome the questionable reliable of observational resea (Schuemie *et al.*, 2018a). This LEGEND document repo the risk of angioedema between new users of lisinopril a hydrochlorothiazide treated for Hypertension.

Worldwide, hypertension stands as a leading cause of m tality, with an increasing prevalence over the last two deca (Forouzanfar *et al.*, 2017). The 2017 American College Cardiology (ACC) / American Heart Association (AHA) cl cal practice guidelines define hypertension based on averag blood pressure (BP) measured in a healthcare setting; syst BP between 130 - 139 mmHg or diastolic BP between 8 89 mmHg characterize stage 1 hypertension, and systolic > 140 mmHg or diastolic BP > 90 mmHg mark stage 2

tures laboratory tests for a subset of the covered lives. This administrative claims database includes a variety of fee-for- service, preferred provider organizations, and capitated health plans. The study period spans from 2000-12-31 to 2018

Exposed:
lisinopril: n = 7546771
hydrochlorothiazide: n = 9385478

Study design. This study follows a retrospective, observa comparative cohort design (Ryan *et al.*, 2013) . We i patients who are first time users of lisinopril or hydroch izide, and who have a diagnosis of hypertension on or p treatment initiation. We require that patients have conti observation in the database for at least 365 days prior to ment initiation. We exclude patients with prior angioe events and less than 1 day at risk. Links to full cohort include concept codes, are provided in the Supporting mation. The outcome of interest is angioedema. We the outcome risk window 1 day after treatment initiatio consider two design choices to define the window end. we end the outcome time-at-risk window at first occ of continuous drug exposure, analogous to an on-trea design and, second, we end the outcome time-at-risk w when the patient is no longer observable in the database ogous to an intent-to-treat design. Continuous drug exp are constructed from the available longitudinal data h sidering sequential prescriptions that have fewer than 3 gap between prescriptions.

Statistical analysis. We conduct our cohort study usi open-source OHDSI CohortMethod R package (Schuemie 2018c) , with large-scale analytics achieved through th clops R package (Suchard *et al.*, 2013) . We use prop scores (PSs) - estimates of treatment exposure probabili ditional on pre-treatment baseline features in the one yea to treatment initiation - to control for potential measure fosdhing and improve balance between the target (lisi and comparator (hydrochloorothiazide) cohorts (Rosen and Rubin, 1983) . We use an expansive PS model th cludes all available patient demographics, drug, conditio procedure covariates generated through the FeatureExtr R package (Schuemie *et al.*, 2018d) instead of a prespecif of investigator-selected confounders. We perform PS stra tion or variable-ratio matching and then estimate compa lisinopril-vs-hydrochlorothiazide hazard ratios (HRs) a Cox proportional hazards model. Detailed covariat methods informations are provided in the Supporting In tion. We present PS and covariate balance metrics to successful confounding control, and provide HR estimat Kaplan-Meier survival plots for the outcome of angioe We additionally estimate HRs for pre-specified subgro evaluate interactions with the treatment effect. For effi reasons, we fit subgroup Cox models using PS stratifi only.

Residual study bias from unmeasured and syste sources can exist in observational studies after controlli measured confounding (Schuemie *et al.*, 2014, 2016) . T mate such residual bias, we conduct negative control ou experiments with 286 negative control outcomes ide through a data-rich algorithm (Voss *et al.*, 2017). We negative control estimates to an empirical null distrib that characterizes the study residual bias and is an imp artifact from which to assess the study design (Schuemie 2018a) . Using the empirical null distribution and syn positive controls (Schuemie *et al.*, 2018b) , we additi

Table 1. Patient cohorts. Target (T) cohort is lisinopril new-users. Comparative (C) cohort is hydrochlorothiazide new-users. We report total number of patients, follow-up time (in years), number of angioedema events, and event incidence rate (IR) per 1,000 patient years (PY) in patient cohorts, as well as the their minimum detectable relative risk (MDRR). Note that the IR does not account for any stratification or matching.

	Patients	
Design	T	C
PS stratification, on-treatment	647,212	286,973
PS stratification, intent-to-treat	648,919	287,953

Fig. 2. Preference score distribution for lisinopril and hydrochlorothiazide ne users. The preference score is a transformation of the propensity score that adju for prevalence differences between populations. A higher overlap indicates t subjects in the two populations are more similar in terms of their predicted probabil of receiving one treatment over the other.

group means before and after PS trimming and stratificati (Table 2). Figure ?? plots StdDiff for all 8843 base-line patie features that serve as input for the PS model. Before strati cation, 118 features have a StdDiff > 0.1. After stratificatio the count is 0.

Outcome assessment. Table 3 details the time to first a gioedema or censoring distributions for patients in the lisin pril and hydrochlorothiazide cohorts. We report in Table estimated HRs comparing lisinopril to hydrochlorothiazi for the on-treatment and intent-to-treat designs with stra fication or matching. Figure 4 plots Kaplan-Meier survivi curves for patients under the intent-to-treat design. To e amine possible subgroup differences in treatment-effect, include Table[tab:subgroups] that reports HR estimates sep rately for children (age < 18), the elderly (age ≥ 65), fema patients, pregnant women, patients with hepatic impairme and patients with renal impairment, using PS stratification

Residual systematic error. In the absence of bias, we expe 95% of negative and positive control estimate 95% confiden intervals to include their presumed HR. In the case of negati controls, the presumed HR = 1. Figure 5 describes the negati and positive control estimates under the on-treatment with P stratification design. Before calibration, negative and positi controls demonstrate poor coverage. After calibration, contro demonstrate poor coverage.

Fig. 5. Evaluation of effect estimation between lisinopril and hydrochlo calibration for each negative and synthetic positive control. The bottom plots th

Conclusions

We find that lisinopril has a higher risk of angioedema compared to hydrochlorothiazide within the population the CCAE represents.

Supporting Information

Here we enumerate the guiding principles of LEGEND provide linking details on study cohorts and design.

LEGEND principles.

1. Evidence will be generated at large-scale.
2. Dissemination of the evidence will not depend on estimated effects.
3. Evidence will be generated by consistently applyin systematic approach across all research questions.
4. Evidence will be generated using a pre-specified ana design.
5. Evidence will be generated using open source soft that is freely available to all.
6. Evidence generation process will be empirically evalu by including control research questions where the effect size is known.
7. Evidence will be generated using best-practices.
8. LEGEND will not be used to evaluate methods.
9. Evidence will be updated on a regular basis.
10. No patient-level data will be shared between sites in network, only aggregated data.

Study cohorts. Please see the LEGEND hyperten Study protocol (https://github.com/OHDSI/Legend/tree/ma Documents) for complete specification of the lisinopril, drochlorothiazide and angioedema cohorts using ATl (http://www.ohdsi.org/web/atlas).

Negative controls. We selected negative controls using a cess similar to that outlined by Voss *et al.* (2017). We construct a list of all conditions that satisfy the follo criteria with respect to all drug exposures in the LEGE hypertension study:

- No Medline abstract where the MeSH terms sugges drug-condition association (Wimmenburg *et al.*, 2015

Table 2. Patient demographics. We report the standardized differ- ence of population means (StdDiff) before and after stratification for selected base-line patient characteristics.

	Before stratification			After stratification		
Characteristic	T (%)	C (%)	StdDiff	T (%)	C (%)	StdDiff
Age group						
10-14	0.1	0.1	0.02	0.1	0.1	0.00
15-19	0.7	0.6	0.01	0.7	0.7	0.00
20-24	1.4	1.6	-0.01	1.5	1.5	0.00
25-29	2.7	3.0	-0.03	2.9	2.9	0.00
30-34	5.2	5.6	-0.04	5.5	5.6	0.00
35-39	8.4	8.9	-0.05	8.8	8.8	0.01
40-44	12.3	13.0	-0.03	12.8	12.8	0.00
45-49	16.2	16.6	-0.02	16.3	16.1	0.00
50-54	18.6	17.6	0.02	18.4	18.3	0.00
55-59	18.0	16.0	0.05	17.4	17.6	-0.01
60-64	15.1	13.1	0.06	14.7	14.7	0.00
65-69	1.8	1.1	0.02	1.2	1.2	0.00
Gender: female	39.2	61.1	-0.47	48.2	48.7	-0.01
Medical history: General						
Acute respiratory disease	24.8	24.3	0.03	25.3	25.6	-0.01
Attention deficit hyperactivity disorder	1.2	1.1	0.02	1.2	1.2	0.00
Chronic liver disease	1.8	1.1	0.04	1.6	1.3	0.03
Chronic obstructive lung disease	1.7	1.6	0.02	1.6	1.7	-0.01
Crohn's disease	0.3	0.2	0.01	0.3	0.3	0.00
Dementia	0.1	0.1	0.00	0.1	0.1	-0.01
Depressive disorder	7.7	8.1	-0.02	7.9	8.2	-0.01
Diabetes mellitus	17.6	4.6	0.43	13.6	13.7	0.00
Gastroesophageal reflux disease	8.0	7.6	0.02	7.9	8.2	-0.01
Gastrointestinal hemorrhage	1.8	1.6	0.00	1.7	1.7	0.00
Human immunodeficiency virus infection	0.2	0.3	-0.01	0.2	0.3	0.00
Hyperlipidemia	26.8	25.0	0.20	32.5	32.5	0.00
Lesion of liver	0.3	0.2	0.01	0.2	0.3	-0.01
Obesity	9.0	9.9	-0.03	9.2	9.5	-0.01
Osteoarthritis	11.4	10.4	0.02	11.2	11.6	-0.01
Pneumonia	1.8	1.4	0.00	1.5	1.6	-0.01
Psoriasis	1.2	0.9	0.02	1.0	1.0	0.00
Renal impairment	1.2	0.5	0.06	0.9	1.0	-0.01
Rheumatoid arthritis	0.8	0.6	0.01	0.8	0.8	-0.01
Schizophrenia	0.1	0.1	0.00	0.1	0.1	0.00
Ulcerative colitis	0.3	0.2	0.01	0.3	0.3	0.00
Urinary tract infectious disease	9.0	8.4	0.02	8.5	8.7	-0.01
Viral hepatitis C	0.4	0.3	0.01	0.4	0.4	0.00
Visual system disorder	15.4	14.9	0.02	15.4	15.4	0.00
Medical history: Cardiovascular disease						
Atrial fibrillation	0.4	0.2	0.03	0.4	0.3	0.00
Cerebrovascular disease	1.7	1.0	0.05	1.5	1.6	-0.01
Coronary artery disease	1.9	1.0	0.08	1.5	1.7	-0.02
Heart disease	8.2	6.6	0.07	7.7	8.2	-0.02
Heart failure	0.4	0.3	0.02	0.3	0.4	0.00
Ischemic heart disease	1.8	0.9	0.06	1.3	1.5	0.00
Peripheral vascular disease	0.2	0.2	0.01	0.2	0.3	-0.01
Pulmonary embolism	0.2	0.2	0.01	0.2	0.3	-0.01
Venous thrombosis	1.0	1.0	0.00	1.0	1.0	0.00
Medical history: Neoplasms						
Hematologic neoplasm	0.5	0.4	0.01	0.5	0.5	-0.01
Malignant lymphoma	0.2	0.2	0.01	0.2	0.2	-0.01
Malignant neoplasm of anorectum	0.1	0.1	0.01	0.1	0.1	-0.01
Malignant neoplastic disease	4.1	3.8	0.02	4.1	4.3	-0.01
Malignant tumor of breast	0.7	1.0	-0.03	0.8	0.8	0.00
Malignant tumor of colon	0.2	0.2	0.01	0.2	0.2	-0.01
Malignant tumor of urinary bladder	0.1	0.1	0.01	0.1	0.1	0.00
Primary malignant neoplasm of prostate	0.5	0.3	0.02	0.4	0.5	-0.01
Medication use						
Antibacterials for systemic use	46.8	30.0	0.34	39.6	39.8	0.00
Antidepressants	19.0	16.3	0.03	18.4	19.0	-0.02
Antidiabetics	6.3	4.0	0.10	6.3	6.5	-0.01
Anti-inflammatory and antirheumatic products	29.3	26.6	0.05	29.0	29.2	0.00
Antineoplastic agents	1.4	1.0	0.03	1.3	1.3	0.00
Antipsychotics	0.4	0.4	0.00	0.4	0.4	0.00
Antithrombotic agents	3.2	2.2	0.06	2.9	3.1	-0.01
Beta blocking agents	0.4	0.4	0.01	0.4	0.4	0.00
Drugs for acid related disorders	13.8	14.0	-0.00	13.9	14.4	-0.01
Drugs for obstructive airway diseases	18.5	20.3	-0.05	19.1	19.4	-0.01
Drugs used in diabetes	15.1	3.1	0.43	11.4	13.0	0.05
Immunosuppressants	1.8	1.6	0.02	1.5	1.6	0.00
Lipid modifying agents	23.8	13.6	0.27	20.6	20.2	0.01
Opioids	15.8	16.0	-0.01	15.7	16.1	-0.01
Psycholeptics	17.7	18.2	-0.01	17.9	18.7	-0.02
Psychoanaleptics, agents used for adhd and nootropics	3.0	3.7	-0.01	3.1	3.1	0.00

References

ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group (2002). "Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: The antihypertensive and lipid-lowering treatment to prevent heart attack trial (ALLHAT)." *Journal of the American Medical Association*, **288**, 2981–2997.

Austin PC (2009). "Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples." *Statistics in medicine*, **28**(25), 3083–3107.

Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH (2016). "A curated and standardized adverse drug event resource to accelerate drug safety research." *Scientific data*, **3**, 160026.

Duke J, Friedlin J, Li X (2013). "Consistency in the safety labeling of bioequivalent medications." *Pharmacoepidemiology and Drug Safety*, **22**(3), 294–301.

Evans S, Waller PC, Davis S (2001). "Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports." *Pharmacoepidemiology and Drug Safety*, **10**(6), 483–486.

Forouzanfar MH, Liu P, Roth GA, Ng M, Biryukov S, Marczak L, Alexander L, Estep K, Abate KH, Akinyemiju TF, *et al.* (2017). "Global burden of hypertension and systolic blood pressure of at least 110 to 115 mm Hg, 1990-2015." *Journal of the American Medical Association*, **317**(2), 165–182.

Hripcsak G, Duke J, Shah N, Reich C, Huser V, Schuemie M, Suchard M, Park R, Wong I, Rijnbeek P, *et al.* (2015). "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers." *Studies in health technology and informatics*, **216**, 574–578.

Overhage J, Ryan P, Reich C, Hartzema A, Stang P (2012). "Validation of a common data model for active safety surveillance research." *Journal of the American Medical Informatics Association*, **19**, 54–60.

Rebouissin DM, Allen NB, Griswold ME, Guallar E, Hong Y, Lackland DT, Miller EPR, Polonsky T, Thompson-Paul AM, Vupputuri S (2017). "Systematic review for the 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines." *Journal of the American College of Cardiology*, p. 24428.

Rosenbaum PR, Rubin DB (1983). "The central role of the propensity score in observational studies for causal effects." *Biometrika*, **70**(1), 41–55.

Ryan PB, Schuemie MJ, Gruber S, Zorych I, Madigan D (2013). "Empirical performance of a new user cohort method: lessons for developing a risk identification and analysis system." *Drug Safety*, **36**(1), 59–72.

Schuemie M, Ryan P, Hripcsak G, Madigan D, Suchard M (2018a). "Improving reproducibility by using high-throughput observational studies with empirical calibration." *Philosophical Transactions of the Royal Society A*, **376**, 20170356.

Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA (2016). "Robust empirical calibration of p-values using observational data." *Statistics in Medicine*, **35**(22), 3883–3888.

Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA (2018b). "Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data." *Proceedings of the National Academy of Sciences*, p. 201708282.

Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D (2014). "Interpreting observational studies: why empirical calibration is needed to correct p-values." *Statistics in medicine*, **33**(2), 209–218.

Schuemie MJ, Suchard MA, Ryan PB (2018c). *CohortMethod: New-user cohort method with large scale propensity and outcome models*. R package version 2.5.1.

Schuemie MJ, Suchard MA, Ryan PB, Reps J (2018d). *FeatureExtraction: Generating Features for a Cohort*. R package version 2.1.5.

Suchard M, Simpson S, Zorych I, Ryan P, Madigan D (2013). "Massive parallelization of serial inference algorithms for a complex generalized linear model." *Transactions on Modeling and Computer Simulation*, **23**,

Different use cases of LEGEND results

- Current use: Hypothesis testing

  question → check LEGEND results

  One hypothesis, no multiple testing.

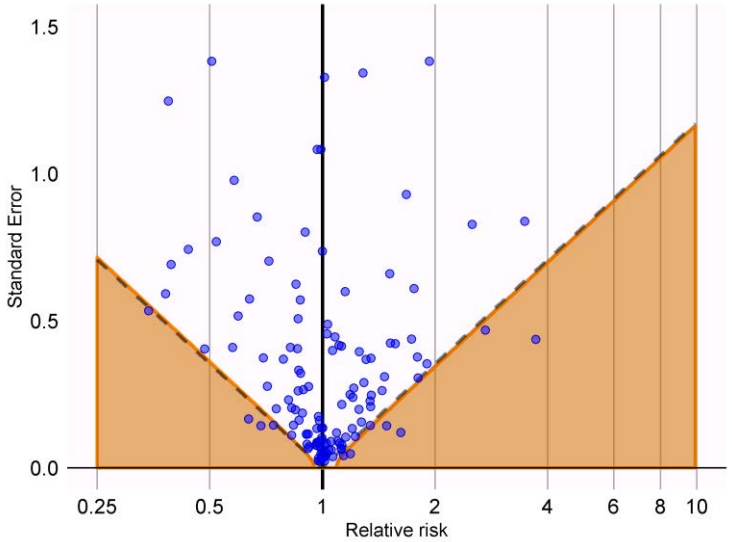
- "Signal detection"

  e.g. rank-order by lower bound of confidence interval

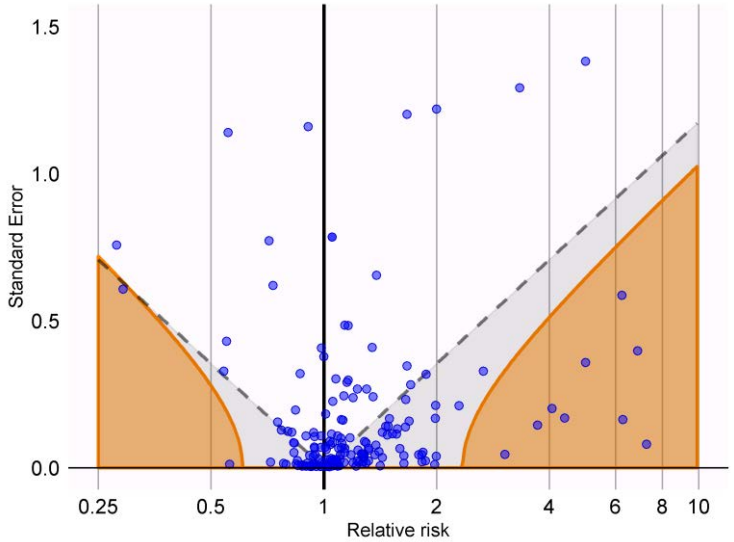
  This requires adjustment for multiple testing!

Same goal: provide best estimate

Estimates for 200 negative controls



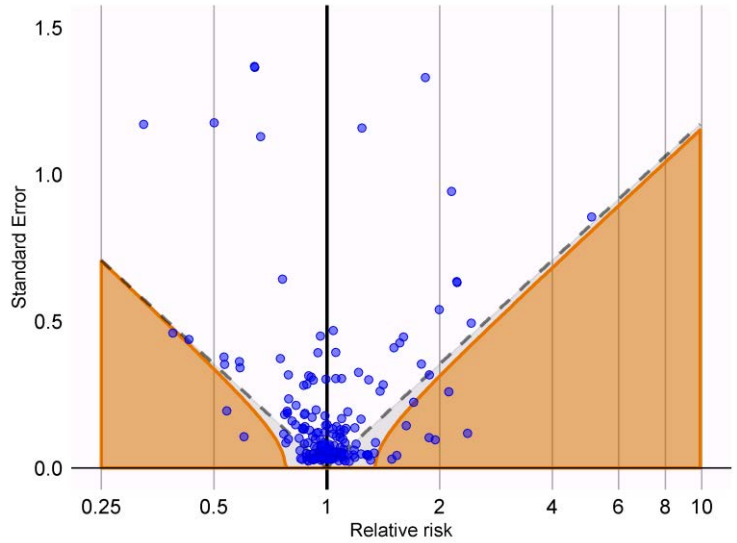
Cohort method using 1-on-1 matching



Nested case-control

Each blue dot is an effect size estimate for a negative control exposure-outcome pair (true relative risk = 1).

Dots below the dashed line have $p < 0.05$.



Multiple Self-controlled Case Series

To evaluate performance, we must decide on the evaluation metrics

- Coverage of the 95% confidence interval
- Type 1 and 2 errors (sensitivity and specificity)
- AUC (Area Under the Receiver Operator Curve)
- MSE (Mean Squared Error)
- Precision

Choice of cutoff (e.g. p=0.05)

Empirical calibration?

Takeaway points

- Goal: to reliably estimate causal effects

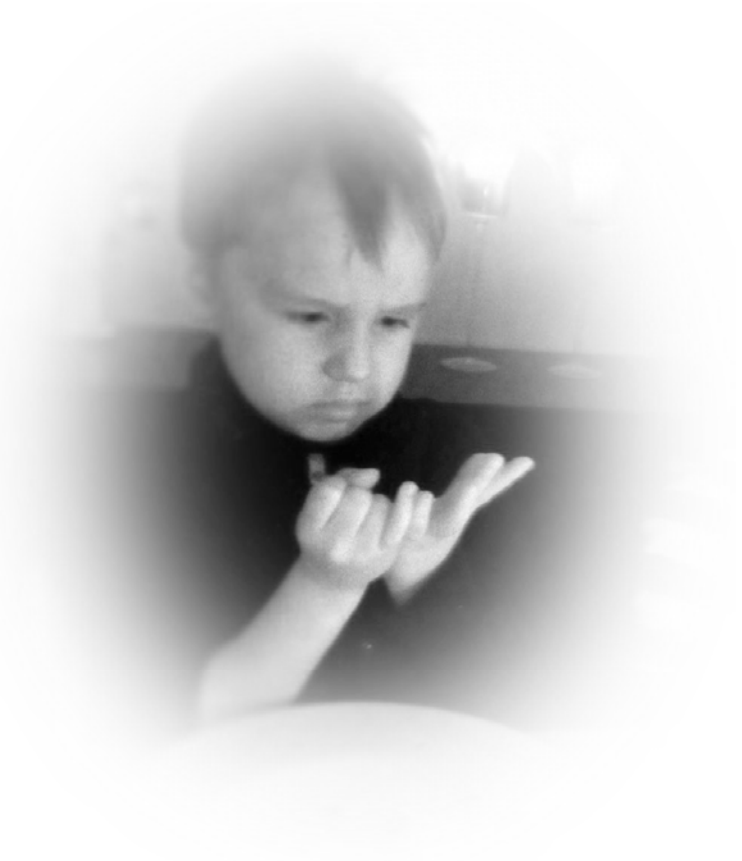
  whether one hypothesis at a time or many hypotheses at a time ("signal detection”)
- Apply best practices, even at large scale
  - Confounding adjustment
  - Proper outcome definitions
  - Sensitivity analyses
  - Study diagnostics
  - Multiple databases
- Always measure operating characteristics using
  - Negative and positive controls
  - Multiple metrics

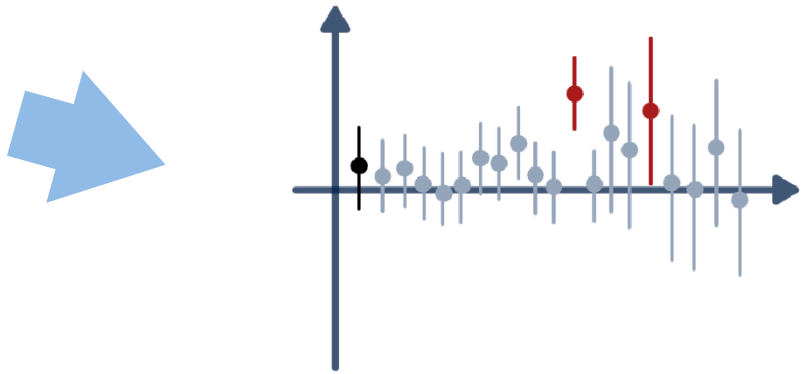
Temporal pattern discovery for signal detection in electronic healthcare data

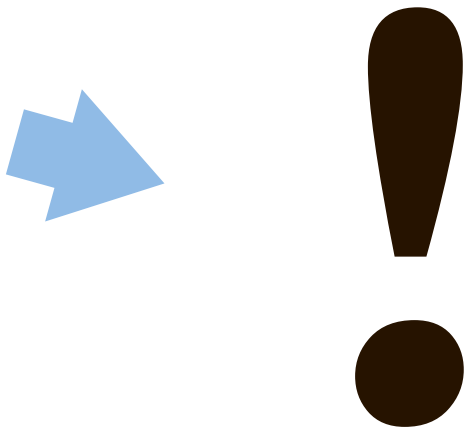
Niklas Norén

Uppsala Monitoring Centre
– Building a global safety culture

Confirmatory study



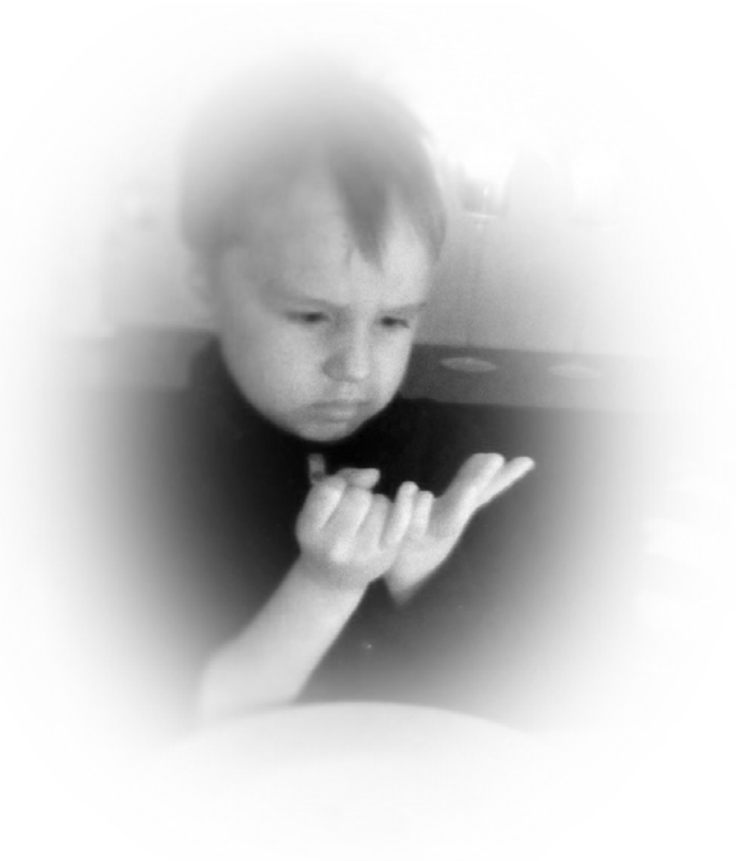


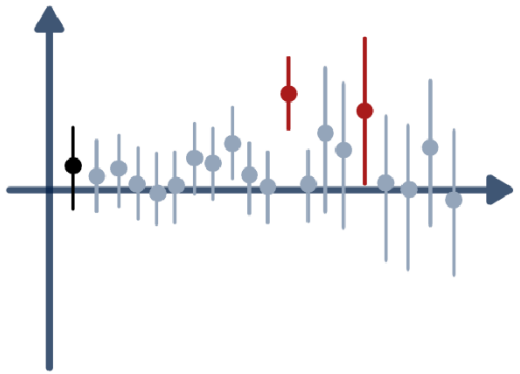


Uppsala Monitoring Centre

– Building a global safety culture

Exploratory study



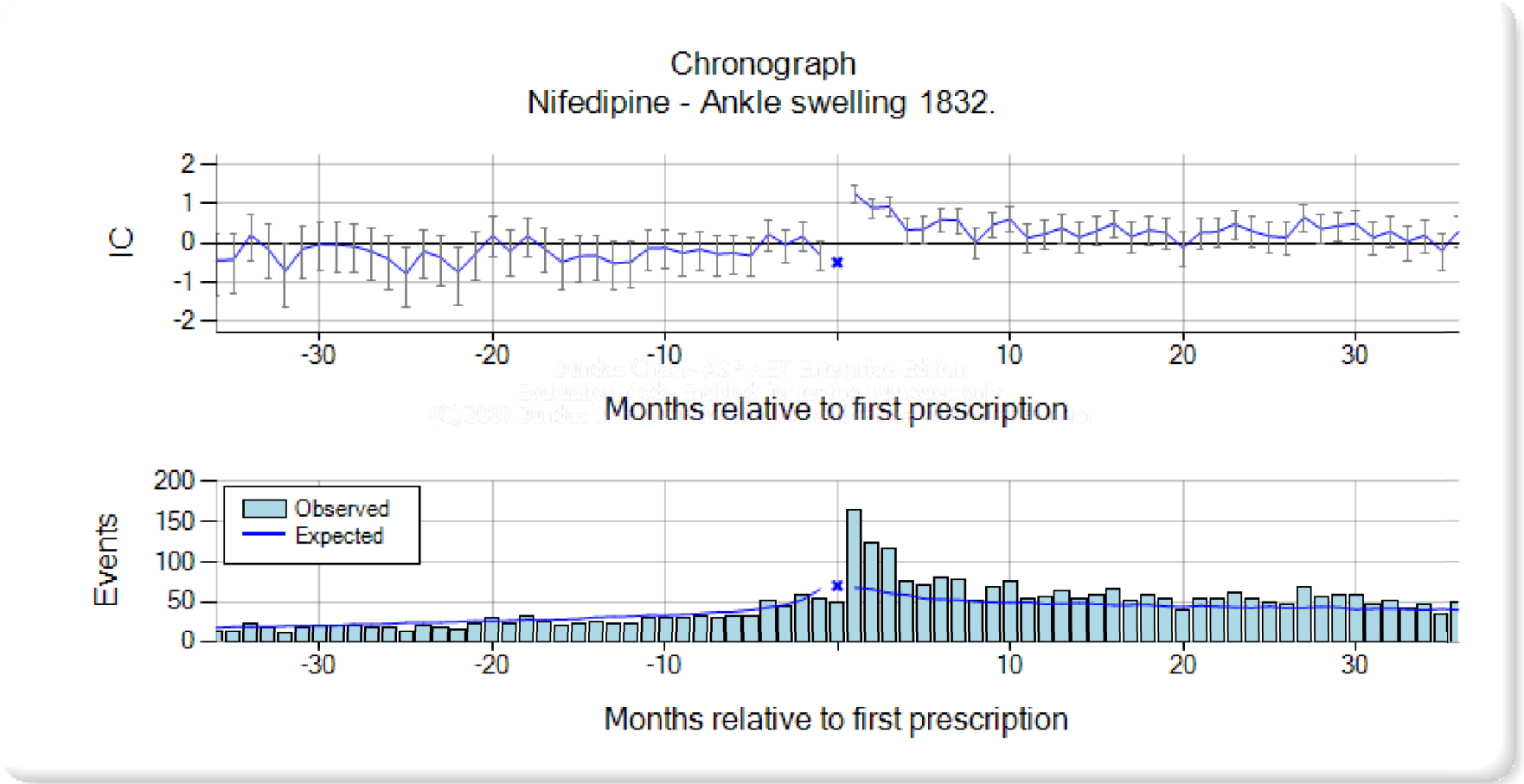


Uppsala Monitoring Centre

– Building a global safety culture

Temporal pattern discovery

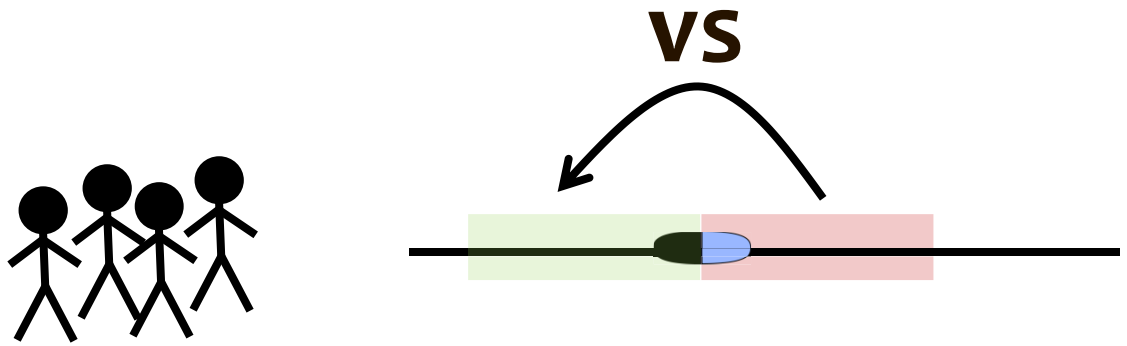
Norén et al. Data Mining and Knowledge Discovery, 2010



Uppsala Monitoring Centre

– Building a global safety culture

Self-control

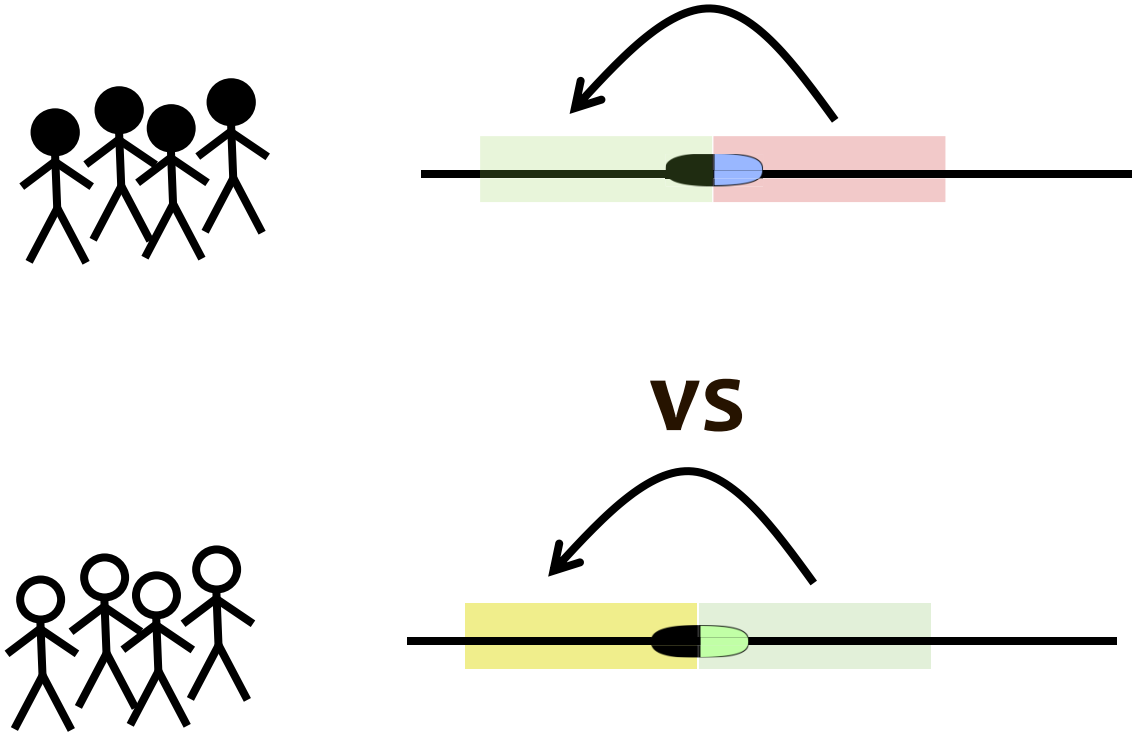


Uppsala Monitoring Centre

– Building a global safety culture

Self-control calibrated by active comparator

Norén et al.
Drug Safety, 2013

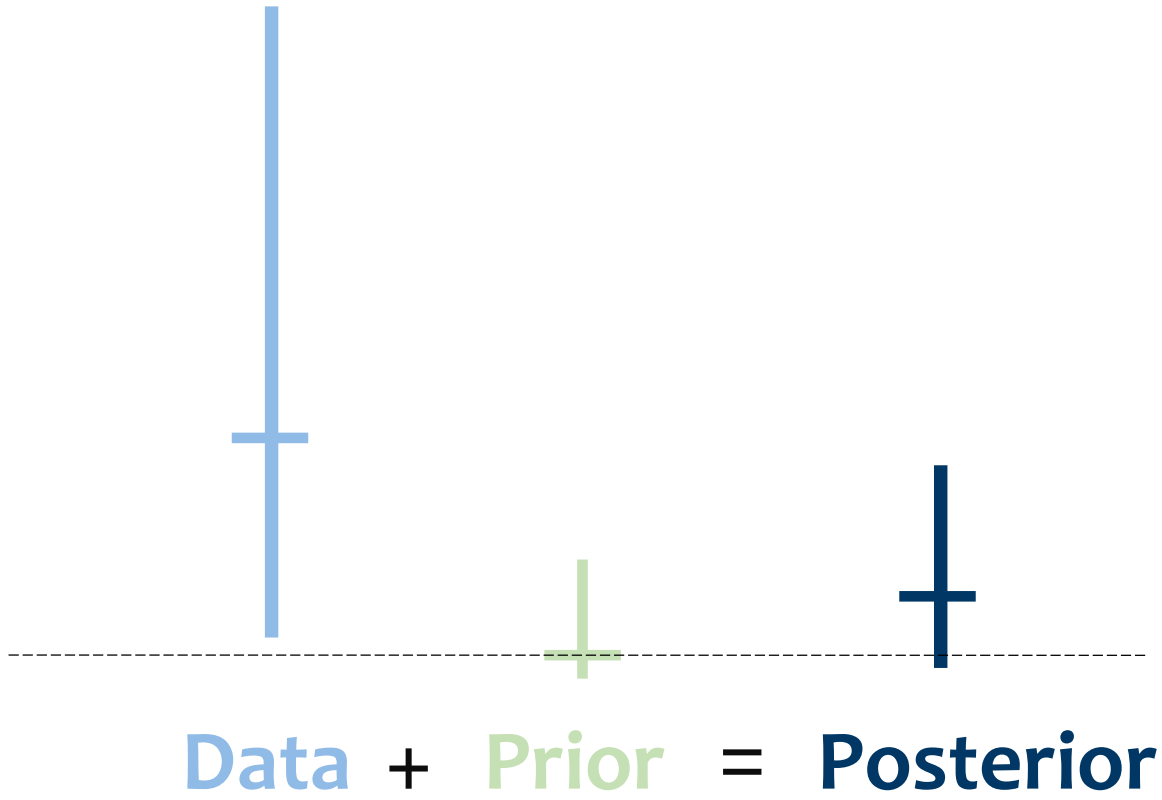


Uppsala Monitoring Centre
– Building a global safety culture

Simple statistical shrinkage

Norén et al
Stat Meth Med Res, 2013

$$\frac{\text{Observed} + 1/2}{\text{Expected} + 1/2}$$

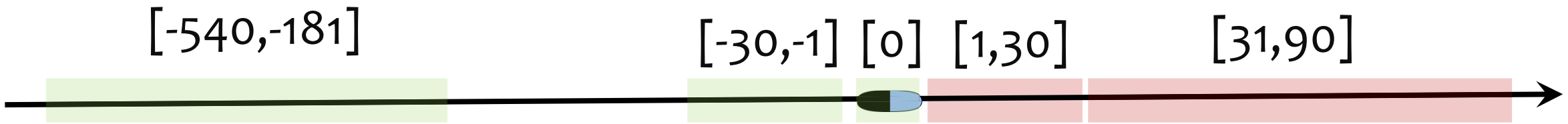


Uppsala Monitoring Centre

– Building a global safety culture

Multiple risk and comparison windows

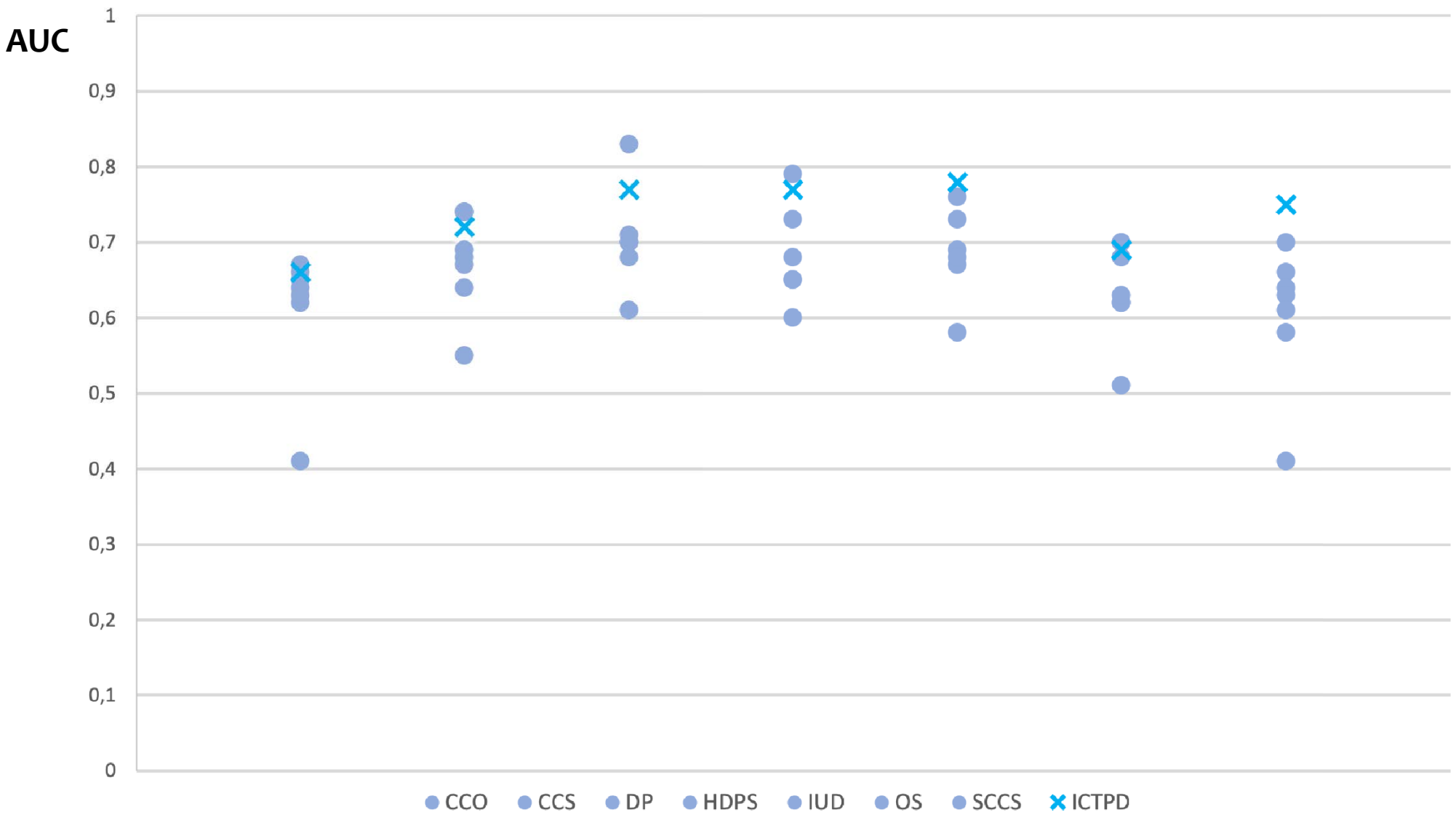
Norén et al.
Drug Safety, 2013



Uppsala Monitoring Centre
– Building a global safety culture

OMOP results (2012 study)

Ryan et al
Drug Safety, 2013



Uppsala Monitoring Centre
– Building a global safety culture

Prospective screening study

Cederholm et al.
Drug Safety, 2015

Drug Saf (2015) 38:87–100
DOI 10.1007/s40264-014-0251-y

ORIGINAL RESEARCH ARTICLE

Structured Assessment for Prospective Identification of Safety Signals in Electronic Medical Records: Evaluation in the Health Improvement Network

S. Cederholm · G. Hill · A. Asiimwe · A. Bate · F. Bhayat · G. Persson Brobert · T. Bergvall · D. Ansell · K. Star · G. N. Norén

Published online: 25 December 2014

Abstract

Background Pharmacovigilance signal detection largely relies on individual case reports, but longitudinal health data are being explored as complementary information sources. Research to date has focused on the ability of epidemiological methods to distinguish established adverse drug reactions (ADRs) from unrelated adverse events.

Objective The aim of this study was to evaluate a process for structured clinical and epidemiological assessment of temporally associated drugs and medical events in electronic medical records.

Methods Pairs of drugs and medical events were selected for review on the basis of their temporal association according to a calibrated self-controlled cohort analysis in The Health Improvement Network. Six assessors trained in pharmacovigilance and/or epidemiology evaluated seven drugs each, with up to 20 medical events per drug. A pre-specified questionnaire considered aspects related to the nature of the temporal pattern, demographic features of the cohort, concomitant medicines, earlier signs and symptoms, and possible confounding by underlying disease. This informed a classification of drug–event pairs as known ADRs, meriting further evaluation, or dismissed.

Results The number of temporally associated medical events per drug ranged from 11 to 307 (median 50) for the 42 selected drugs. Out of the 509 relevant drug–event combinations subjected to the assessment, 127 (25 %) were classified as known ADRs. Ninety-one (24 %) of the remaining pairs were classified as potential signals meriting further evaluation and 291 (76 %) were dismissed. Suggestive temporal patterns and lack of clear alternative explanations were the most common reasons that drug–event pairs were classified as meriting further evaluation. Earlier signs and symptoms and confounding by the underlying disease were the most common reasons that drug–event pairs were dismissed.

Conclusions Exploratory analysis of electronic medical records can detect important potential safety signals. However, effective signal detection requires that statistical signal detection be combined with clinical and epidemiological review to achieve an acceptable false positive rate.

Key Points

Exploratory analysis of electronic medical records can detect important potential safety signals.

To achieve an acceptable false positive rate, statistical signal detection should be combined with clinical and epidemiological review.

Such review also requires a deep understanding of the analytical methods employed, and insight into data collection and medical practice in the setting at hand.

S. Cederholm · G. Hill · T. Bergvall · K. Star · G. N. Norén (✉)
Uppsala Monitoring Centre, WHO Collaborating Centre for International Drug Monitoring, Box 1051, SE-75140 Uppsala, Sweden
e-mail: niklas.noren@who-umc.org

A. Asiimwe
Eli Lilly UK, Surrey, UK

A. Bate
Pfizer UK, Surrey, UK

F. Bhayat
Takeda (TGRD), London, UK

G. Persson Brobert
Bayer Pharma AG, Berlin, Germany

D. Ansell
Cegedim Strategic Data Medical Research, London, UK

Sibutramine

Nifedipine

Oedema

Flushing

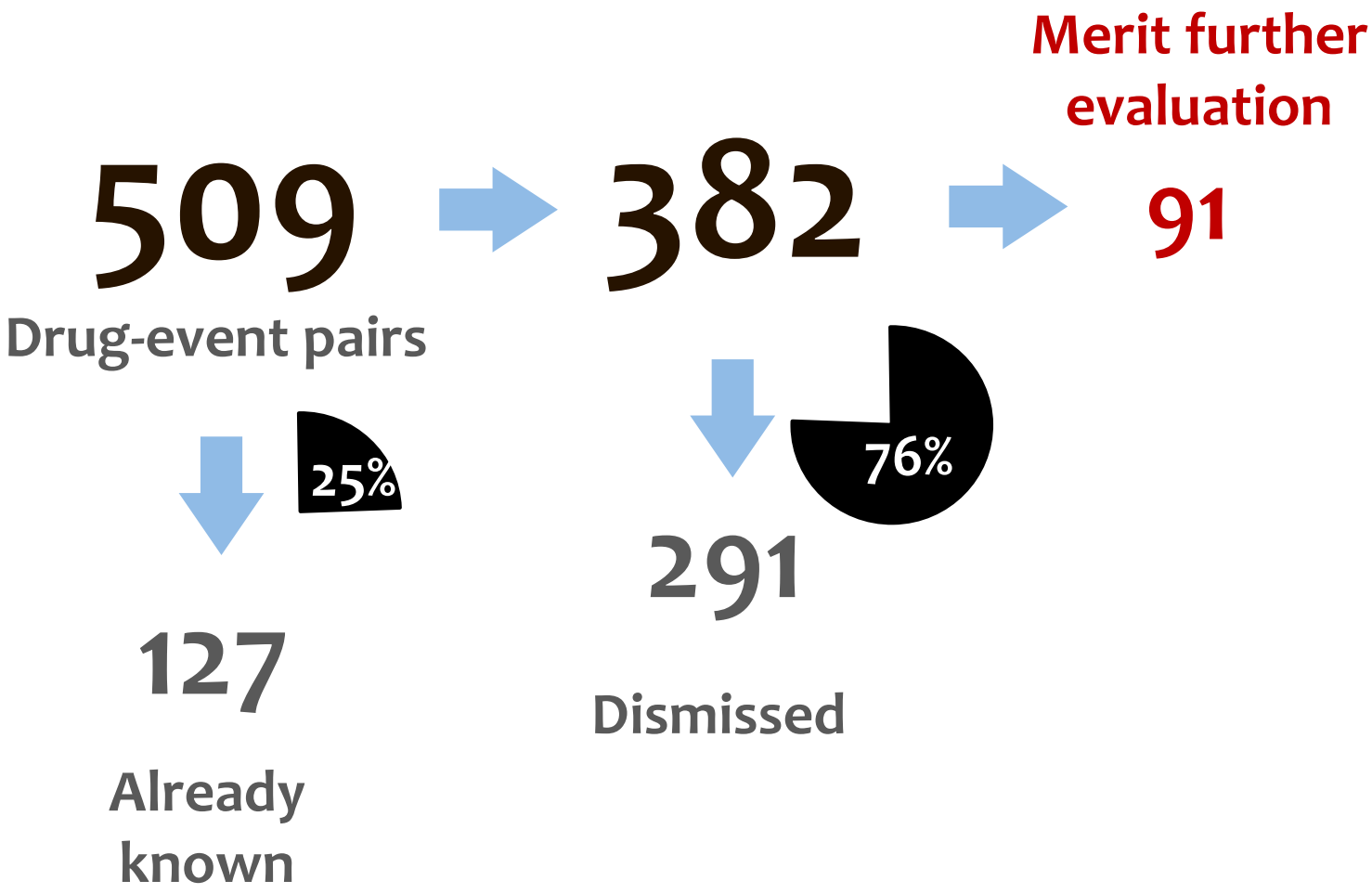
6 assessors x **7** drugs per assessor x **20** events per drug

Uppsala Monitoring Centre

– Building a global safety culture

Prospective screening study

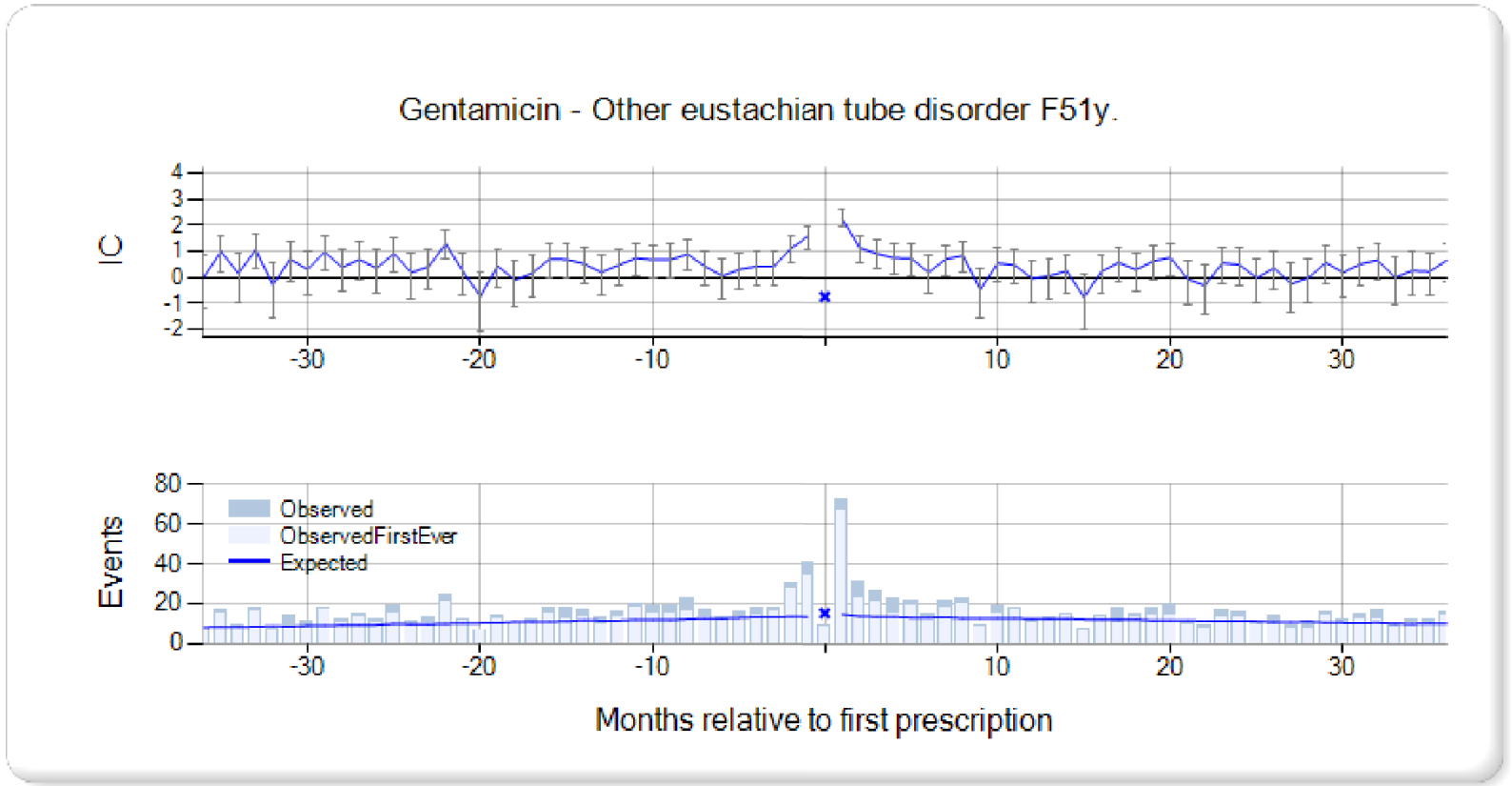
Cederholm et al.
Drug Safety, 2015



Uppsala Monitoring Centre
– Building a global safety culture

Confounding by underlying disease

Cederholm et al.
Drug Safety, 2015

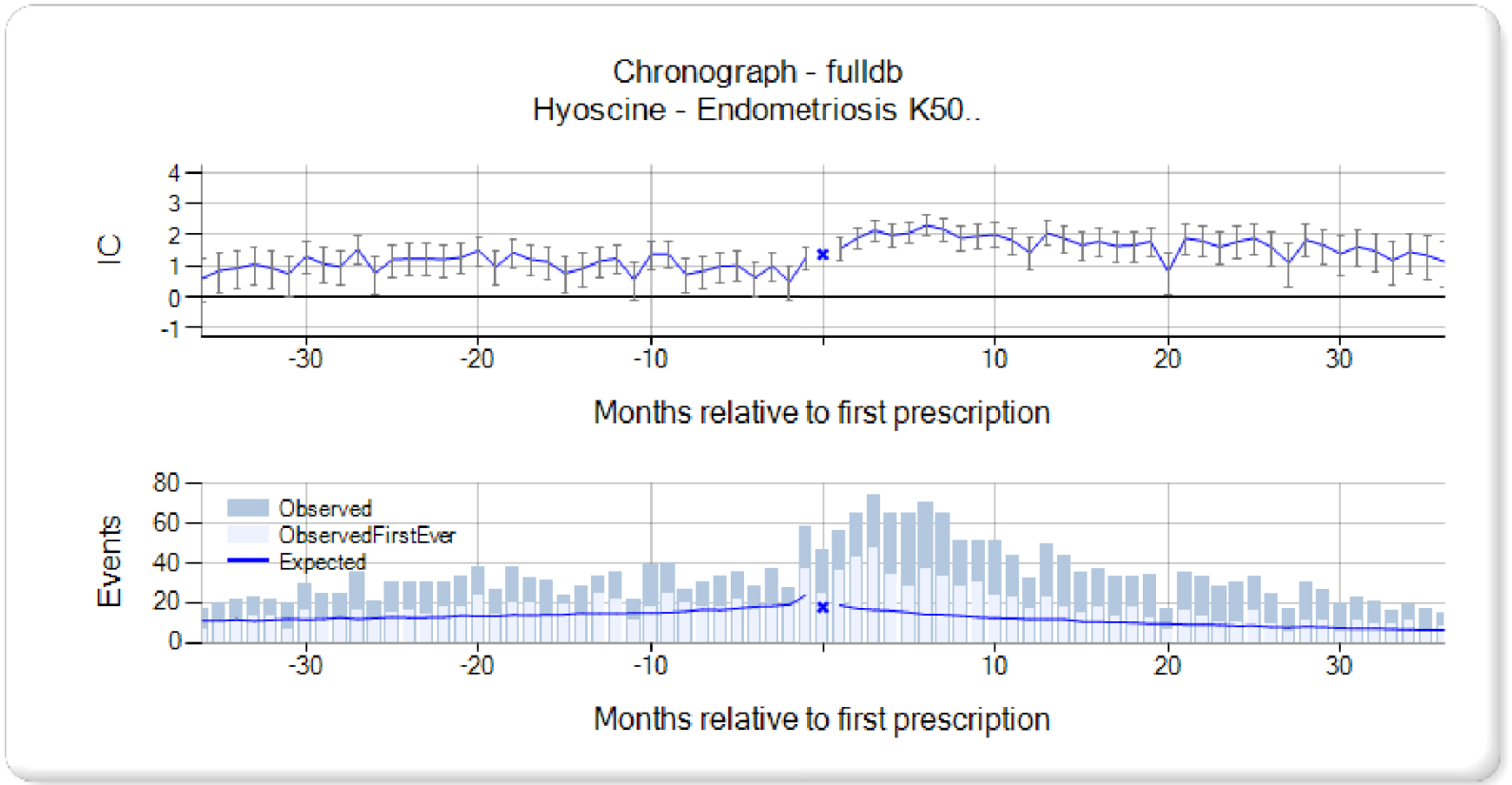


Uppsala Monitoring Centre

– Building a global safety culture

Protopathic bias?

Cederholm et al.
Drug Safety, 2015



Uppsala Monitoring Centre

– Building a global safety culture

References

1. Norén GN, Hopstadius J, Bate A, Star K, Edwards IR. **Temporal pattern discovery in longitudinal electronic patient records**. *Data Mining and Knowledge Discovery*, 2010. **20**(3):361-387.
2. Norén GN, Bergvall T, Ryan P, Juhlin K, Schuemie MJ, Madigan D. **Empirical Performance of the Calibrated Self-Controlled Cohort Analysis within Temporal Pattern Discovery: Lessons for Developing a Risk Identification and Analysis System**. *Drug Safety* 2013; **36**(S1):107-121.
3. Norén GN, Hopstadius J, Bate A. **Shrinkage observed-to-expected ratios for robust and transparent large-scale pattern discovery**. *Statistical Methods in Medical Research* 2013; 22(1):57-69.
4. Ryan PB, Stang PE, Overhage JM, Suchard MA, Hartzema AG, DuMouchel W, Reich CG, Schuemie MJ, Madigan D. **A Comparison of the Empirical Performance of Methods for a Risk Identification System**. *Drug Safety* 2013; **36**(S1):143-158.
5. Cederholm S, Hill G, Asiimwe A, Bate A, Bhayat F, Persson Brobert G, Bergvall T, Ansell D, Star K, Norén GN. **Structured assessment for prospective identification of safety signals in electronic medical records: evaluation in The Health Improvement Network**. *Drug Safety*, 2015; **38**(1):87-100.

Uppsala Monitoring Centre

– Building a global safety culture