# TITLE: ENHANCING DRUG SAFETY THROUGH ACTIVE SURVEILLANCE OF OBSERVATIONAL HEALTHCARE DATA

Patrick B. Ryan

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Division of Pharmaceutical Outcomes and Policy in the UNC Eshelman School of Pharmacy

Chapel Hill
2011

Approved by:

Richard A. Hansen, PhD

Joel F. Farley, PhD

Michael D. Murray, PharmD, MPH

Til Stürmer, MD, MPH

J. Marc Overhage, MD, PhD

# ABSTRACT

PATRICK B. RYAN: Enhancing Drug Safety Through Active Surveillance of Observational Healthcare Data
(Under the direction of Dr. Richard A. Hansen)

Drug safety continues to be a major public health concern in the United States, with adverse drug reactions ranking as the 4th to 6th leading cause of death, and resulting in health care costs of $3.6 billion annually. Recent media attention and public scrutiny of high-profile drug safety issues have increased visibility and skepticism of the effectiveness of the current post-approval safety surveillance processes. Current proposals suggest establishing a national active drug safety surveillance system that leverages observational data, including administrative claims and electronic health records, to monitor and evaluate potential safety issues of medicines. However, the development and evaluation of appropriate strategies for systematic analysis of observational data have not yet been studied.

This study introduces a novel exploratory analysis approach (Comparator-Adjusted Safety Surveillance or COMPASS) to identify drug-related adverse events in automated healthcare data. The aims of the study were: 1) to characterize the performance of COMPASS in identifying known safety issues associated with ACE inhibitor exposure within an administrative claims database; 2) to evaluate consistency of COMPASS estimates across a network of disparate databases; and 3) to explore differential effects across ingredients within ACE inhibitor class.

COMPASS was observed to have improved accuracy to three other methods under consideration for an active surveillance system: observational screening, disproportionality analysis, and self-controlled case series. COMPASS performance was consistently strong within 5 different databases, though important differences in outcome estimates across the sources highlighted the substantial heterogeneity which makes pooling estimates challenging. The comparative safety analysis of products within the ACE inhibitor class provided evidence of similar risk profiles across an array of different outcomes, and raised questions about the product labeling differences and how observational studies should complement existing evidence as part of a broader safety assessment strategy.

The results of this study should inform decisions about the appropriateness and utility of analyzing observational data as part of an active drug safety surveillance process. An improved surveillance system would enable a more comprehensive and timelier understanding of the safety of medicines. Such information supports patients and providers in therapeutic decision-making to minimize risks and improve the quality of care.

# ACKNOWLEDGEMENTS

The achievement of one rests on the shoulders of many. This milestone toward my personal development is really a testament to the ongoing support of everyone around me along the way.

First, I want to thank my mother and father for their unending support and the providing me with belief that all goals are achievable. Thank you for instilling me a conviction that work ethic provides an opportunity to make meaningful and lasting contributions to mankind.

Thank you to all my colleagues at GlaxoSmithKline, Johnson & Johnson, and across Observational Medical Outcomes Partnership. My professional collaborations provided me with the intellectual curiosity to initiate my academic research, but completing this project would not have been achievable without the support and freedom to pursue my professional and scholarly ambitions simultaneously. Thanks to Alan, Paul, and David for giving me the motivation to challenge past practices and to look for better solutions. Thanks to my classmates and faculty at UNC Eshelman School of Pharmacy for allowing me to become part of a community of great intellectual stimulation and collegiality.

Thanks to Rick, for encouraging me to chase my passion, no matter how unconventional the path. I will always appreciate your willingness to support me in my pursuits at every stage along the process, from taking your course as a non-degree-seeking student through joining the program as a “part-time” experiment to sheparding me through the dissertation from afar. Also, thank you to Joel, Til, Mick and Marc for your expert counsel, persistent support, and remarkable patience. Your insight helped me to think more critically about how to help advance the science of pharmacoepidemiology.

Most importantly, thank you to my best friend, my soulmate, my wife Holly. You give me inspiration to reach for my educational ambitions and motivation to aspire to the highest standards. You provided more patience and support than anyone deserves. Your love and encouragement throughout is the main reason why I’ve made it to this point. I will be forever thankful for taking our academic journeys together, and look forward to where the rest of our life journeys take us from here.

# TABLE OF CONTENTS

LIST OF TABLES ........ x

LIST OF ABBREVIATIONS ........ xiii

CHAPTER ONE: INTRODUCTION ........ 1

1.1 Overview ........ 1

1.1 Specific Aims ........ 3

1.3 Importance of Proposed Research Plan ........ 7

CHAPTER TWO: BACKGROUND AND SIGNIFICANCE ........ 9

2.1 The Increasing Importance of Drug Safety in the Quality of Healthcare ........ 9

2.2 History of FDA’s response to drug safety ........ 13

2.3 Approaches for evaluating drug safety issues ........ 14

2.4 Approaches for identifying potential drug safety issues ........ 29

2.5 An integrated active surveillance system within a causal inference framework ........ 39

CHAPTER THREE: METHODS ........ 44

3.1 Overview ........ 44

3.2 COMPASS ........ 45

3.2.1 COMPASS comparator selection ........ 47

3.2.2 COMPASS cohort restrictions and adjustments ........ 52

3.2.3 COMPASS risk windows ........ 60

3.2.4 COMPASS prioritization score ........ 63

3.2.5 COMPASS summary ........ 66

3.3 Data Sources ........ 68

3.4 Experimental design ........ 80

3.5 Performance measures ........ 102

3.5 Data analysis by Aim ........ 106

CHAPTER FOUR: MANUSCRIPT 1: “Systematic identification of drug safety issues in administrative claims data: Performance of hypothesis generation methods for active surveillance” ........ 122

Abstract ........ 122

Background ........ 123

Materials and Methods ........ 125

Results ........ 131

Discussion ........ 135

References ........ 140

Tables and Figures ........ 140

CHAPTER FIVE: MANUSCRIPT 2: “Integrating active drug safety surveillance analyses across a network of observational healthcare databases” ........ 149

Abstract ........ 149

Introduction ........ 151

Methods ........ 153

Results ........ 156

References ........ 166

Tables and Figures ........ 166

CHAPTER SIX: MANUSCRIPT 3: "Comparative safety of ACE inhibitors: Evaluating an active surveillance framework" ........ 172

Abstract ........ 172

Introduction ........ 173

Methods ........ 174

Results ........ 177

Discussion ........ 181

References ........ 193

Tables and Figures ........ 188

CHAPTER SEVEN: CONCLUSION AND DISCUSSION ........ 193

7.1. Motivation for study ........ 197

7.2. Review of study results ........ 200

7.3. Lessons through the evolution of the research program ........ 207

7.4. Limitations of the COMPASS method ........ 210

7.5. Limitations of observational data ........ 212

7.6. Limitations of the COMPASS experiments ........ 214

7.7. Contributions to the field ........ 216

WORKS CITED ........ 219

# LIST OF TABLES

Table 1: Source population characteristics ........................................................................ 70

Table 2: ACE inhibitor use across databases ...................................................................... 86

Table 3: Indication covariates identified by COMPASS for each ACE inhibitor ................. 87

Table 4: Comparator drugs selected by COMPASS for each ACE inhibitor ......................... 88

Table 5: Contraindications used as restriction criteria by COMPASS for each ACE inhibitor ............................................................................................................... 91

Table 6: Labeled events identified in SPLs by ingredient ..................................................... 95

Table 7: ACE Inhibitor 'true positive' reference set .............................................................. 97

Table 8: ACE Inhibitor negative controls, by prevalence .................................................... 100

Table 9: Adverse events to explore across ACE inhibitor ingredients ................................ 120

Manuscript 1 Table 1: Operating characteristics of the four methods at alpha=0.05 .......... 140

Manuscript 1 Table 2: Performance measures of the four methods .................................... 140

Manuscript 1 Table 3: COMPASS propensity score balance effects .................................. 141

Manuscript 2 Table 1: Data source characteristics ............................................................. 166

Manuscript 2 Table 2: Operating characteristics of COMPASS across data sources and within composite summaries ........................................................................... 167

Manuscript 3 Table 1: Cohorts, baseline characteristics, indications .................................. 188

Manuscript 3 Table 2: Event rates by ingredient ................................................................ 191

## LIST OF FIGURES

Figure 1: Sources of risk for medical products ........ 10

Figure 2: Generation of health care utilization databases and potential sources of errors and bias[14] ........ 22

Figure 3: Population restrictions to approximate clinical trials in observational studies[88] ..... 24

Figure 4: Conceptual framework for active surveillance ........ 40

Figure 5: COMPASS conceptual model ........ 46

Figure 6: Attributes of medical products used in COMPASS automated heuristics ........ 49

Figure 7: Example attributes for lisinopril ........ 50

Figure 8: COMPASS automated comparator selection heuristic ........ 52

Figure 9: COMPASS pre-exposure design considerations ........ 53

Figure 10: COMPASS automated design refinement process ........ 57

Figure 11: COMPASS alternative risk windows ........ 62

Figure 12: COMPASS prioritization across risk windows ........ 63

Figure 13: COMPASS prioritization across outcomes ........ 66

Figure 14: OMOP Common Data Model conceptual schema ........ 72

Figure 15: Drug era construction ........ 75

Figure 16:  Condition era construction ........ 78

Figure 17: Performance measures for 2x2 contingency table ........ 103

Manuscript 1 Figure 1: COMPASS conceptual model ........ 142

Manuscript 1 Figure 2: Sensitivity and specificity of four methods at alpha=0.05 and 0.001. ........ 143

Manuscript 1 Figure 3: Receiver operating characteristics curves for each of the four methods. ........ 144

Manuscript 1 Figure 4: Estimates for 'label event' conditions across the four methods, ordered by estimate from COMPASS. ........ 145

Manuscript 2 Figure 1: Forest plots of effect estimates for 42 ACE inhibitor labeled events........ 168

Manuscript 2 Figure 2: Heterogeneity across drug-outcome pair ........ 169

Manuscript 3 Figure 1: Population size by propensity score strata ........ 189

Manuscript 3 Figure 2: Impact of propensity score adjustment on measured covariates..... 190

Manuscript 3 Figure 3: Effects estimates by outcome........ 192

# LIST OF ABBREVIATIONS

| | |
|---|---|
| ACE | Angiotensin Converting Enzyme |
| ADR | Adverse Drug Reaction |
| AERS | Adverse Event Reporting System |
| AUC | Area under Receiver Operator Characteristic curve (ROC analysis) |
| CCAE | Thomson Reuters MarketScan Commercial Claims and Encounters |
| COMPASS | COMParator-Adjusted Safety Surveillance |
| CPT-4 | Current Procedural Terminology, 4th edition |
| DP | Disproportionality Analysis |
| FDA | Food and Drug Administration |
| FPR | False Positive Rate (1-specificity) |
| GAO | General Accounting Office |
| HCPCS | Healthcare Common Procedure Coding System |
| IC | Information Component |
| ICD9 | International Classification of Diseases - Clinical Modification 9 |
| INPC | Indiana Network for Patient Care |
| MAP | Mean Average Precision |
| MDCD | MarketScan Medicaid Multi-State Database |
| MDCR | MarketScan Medicare Supplemental and Coordination of Benefits Database |
| MedDRA | Medical Dictionary for Regulatory Activities |

| | |
|---|---|
| MGPS | Multi-item Gamma Poisson Shrinker |
| MSLR | MarketScan Lab Database |
| NDC | National Drug Code |
| OMOP | Observational Medical Outcomes Partnership |
| OS | Observational Screening |
| p@k | Precision-at-k (accuracy measure of precision at the k-th ranked score) |
| PT | MedDRA preferred term |
| r@fp | Recall-at-false positive rate (accuracy measure of sensitivity at a defined false positive threshold) |
| SNOMED | Systematized Nomenclature of Medicine -- Clinical Terms |
| SPL | Structured Product Label |
| SPLICER | Structured Product Label Information Coder and Extractor |
| USCCS | Univariate self-controlled case series |

# CHAPTER ONE: INTRODUCTION

## 1.1 Overview

Drug safety continues to be a major public health concern in the United States. In order for patients and health care providers to make appropriate therapeutic decisions, they need to be informed of the potential benefits and harms of alternative treatment options. While the efficacy of prescription medicines is generally well-characterized from the series of randomized clinical trials conducted during drug development, the safety profile of medicines is often less certain and poorer understood[1]. Research suggests that drug safety information is the highest information priority for patients, and that the perception of side effects is influential in many patients' decisions about taking a medicine[2]. This patient focus is well-justified. Lazarou et al estimated that, in 1994, between 76,000 and 137,000 hospital patients died from an adverse drug reaction (ADR), ranking adverse drug reactions as the fourth to sixth leading cause of death[3], and resulting in health care costs of $3.6 billion annually[4].

The frequency of new safety information being brought to light following regulatory approval is quite striking. A study by the US General Accounting Office (GAO) concluded that 51% of all approved drugs had at least one serious adverse drug reaction that was not recognized during the approval process[5]. A revised estimate from 1994-1997 showed that 30% of products required significant label changes following introduction[6]. Nearly 20 million patients in the United States took at least one of the five drugs withdrawn from the market between September 1997 and September 1998. Seven drugs approved since 1993 and subsequently withdrawn from the market have been reported as possibly contributing to 1002 deaths[7]. It could be speculated that at least some of these negative outcomes could have been averted had the full safety profile been understood at the time of therapeutic decision-making.

Traditional methods of drug safety surveillance involve literature searching and case-by-case analysis of spontaneous adverse event reports, as well as crude frequency counts and calculation of reporting rates[6]. Statistical data mining algorithms are becoming increasingly popular supplementary tools for safety reviewers[8]. Currently, the FDA conducts spontaneous data mining by applying the Multi-item Gamma Poisson Shrinker (MGPS) method to the Adverse Event Reporting System (AERS) database[9, 10]. Many groups have recognized the significant limitations in the current system. As part of the FDA Amendment Act of 2007, Congress mandated the use of observational data (including administrative claims and electronic health records) as part of an active drug safety surveillance system that would supplement the current practice[11].

It is expected that a national active surveillance system will consist of several interlocked processes, including signal detection, signal strengthening, signal validation, and hypothesis testing in a formal pharmacoepidemiologic study[12]. While these observational data sources have been actively studied for pharmacoepidemiologic evaluation studies[13, 14], appropriate statistical methods for screening observational data to generate and triage hypotheses about potential drug effects are nascent and have not yet been rigorously explored across a network of disparate data sources. An outstanding research need is to characterize how well these tools identify true drug-event associations and minimize the number of false positive findings.

We have developed a novel exploratory analysis approach to observational data for active surveillance. The method, called COMParator-Adjusted Safety Surveillance (COMPASS), is a statistical algorithm that estimates adjusted rate differences and relative risks for all outcomes of interest for a given medical product through propensity score stratification across exposed and unexposed cohorts. COMPASS applies an automated heuristic for defining a comparator group based the indication of the medical product, and provides multivariate adjustment based on key influents of risk, including person demographics, comorbidity, and health service utilization. COMPASS is not intended to be a final solution for active surveillance, but instead a first-pass screening tool to serve as a potential guide for identifying and prioritizing potential drug effects that may warrant further evaluation. A goal of this research is to empirically evaluate the behavior of COMPASS to inform its appropriate use within an active surveillance network.

### 1.1 Specific Aims

To study the performance of the novel method, drugs within the Angiotensin Converting Enzyme (ACE) Inhibitor class were explored. ACE Inhibitors provide a solid basis for methodological research because the class represents a large set of mature products that are actively used in the broad population. The safety profile of ACE inhibitors is thought to be well-characterized, including a broad set of known safety issues that span the continuum from common, nuisance effects, such as cough, to rare and more serious events, like angioedema and renal dysfunction. An analysis of the product labels within the ACE inhibitor class identified 50 distinct adverse events listed on the majority of products, 12 of which were highlighted in Boxed Warnings or in Warnings and Precautions sections.

The specific aims of the study were:

**Aim 1: Characterize the performance of COMPASS in identifying known safety issues associated with ACE inhibitor exposure within an administrative claims database**

This aim studied how COMPASS performs in the Thomson Reuters MarketScan Commercial Claims and Encounters (CCAE), a large administrative claims database containing 59 million privately insured lives. CCAE provides patient-level de-identified data from inpatient and outpatient visits and pharmacy claims of multiple insurance plans. CCAE contains 3 million persons with at least one prescription dispensing record for an ACE inhibitor.

COMPASS was applied to the ACE Inhibitor drug class to generate estimates of outcome relationships for a defined set of potential adverse events. These outcomes included both the known associations previously characterized in the product label as well as a sample of 'negative control' conditions for which there is no evidence of drug-related effects. Descriptive statistics summarized the distribution of the estimates and patterns across attributes of the conditions, such as background prevalence rate, confidence in association, and expected degree of confounding.

The objective of a hypothesis-generating tool is to accurately distinguish between true and false relationships. The performance of COMPASS was characterized through multiple measures of accuracy, including area under receiver operator curve[15]. These measures were compared to those from three alternative methods for active surveillance signal generation: disproportionality analysis, as adapted from spontaneous data mining[16, 17]; observational screening, an unadjusted cohort-based design[18]; and, univariate self-controlled case series[19, 20].

**Aim 2: Evaluate consistency of COMPASS estimates across a network of disparate databases**

An active surveillance network is likely to comprise multiple data sources, as it is recognized that there is currently no single US-based source that can be expected to satisfy all requirements of allowing investigation of all medical products for all potential adverse events and across all populations of interest. However, there is little research to inform the expected behavior of active surveillance analysis methods when applied to disparate databases, or the potential benefits of integrating estimates across sources to improve method performance.

This aim conducted the COMPASS analysis for ACE inhibitors across five databases. In addition to CCAE, the method was applied to the MarketScan Lab Database (MSLR), MarketScan Medicaid Multi-State Database (MDCD), MarketScan Medicare Supplemental and Coordination of Benefits Database (MDCR), and the GE Centricity electronic health record (GE). MSLR contains 1.5 million persons representing a largely privately-insured population, with administrative claims from inpatient, outpatient, and pharmacy services supplemented by laboratory results. MDCD provides administrative claims data for 11 million Medicaid enrollees from multiple states. MDCR captures administrative claims for 5 million retirees with Medicare supplemental insurance paid for by employers, including services provided under the Medicare-covered payment, employer-paid portion, and any out-of-pocket expenses. GE contains patient-level data for 11 million persons captured at the point of care from a consortium of providers using the GE Centricity electronic health record system in their outpatient and specialty practices.

$I^2$ statistics were computed to assess the heterogeneity in COMPASS estimates across data sources. Accuracy measures from each source were also compared to assess the reliability in performance. In addition, we explored the use of fixed and random effects meta-analysis[21] to produce composite estimates. We then evaluated the relative performance of the pooled estimate in predicting drug safety issues as compared to source-specific performance to assess the potential advantages of a network-based approach to active surveillance.

**Aim 3: Explore differential effects across ingredients within ACE inhibitor class**

The general consensus within the clinical community is that all ACE inhibitors have similar safety profiles[22]. However, examination of the product labels suggests differences in which adverse events have been reported. Further, there is little information to assess the relative effect size of adverse events across products. This aim applied COMPASS to seven medical products within the class (lisinopril, moexipril, quinapril, ramipril, benazepril, captopril, and enalapril), to determine whether meaningful differences are observed within observational databases.

Among the true relationships, six events (asthma, back pain, bronchospasm, flushing, epistaxis, and tinnitus) are differentially listed on the product labels, indicating the potential to observe different rates among the conditions between products. In addition, 17 events (including abdominal pain, cough, constipation, leucopenia, renal impairment, pruritis, and thrombocytopenia) are consistently recorded across the ingredient labels but no quantitative evidence is provided to compare the strength of the association. COMPASS estimates were summarized to evaluate the relative consistency in risks across individual products, and explore whether differences in product labeling reflect true observed clinical differences between these medicines.

## 1.3 Importance of Proposed Research Plan

An improved drug surveillance system would enable a more comprehensive and timelier understanding of the safety of medicines. Such information will support patients and providers in therapeutic decision-making to minimize risks and improve the quality of care. The results of this study will inform decisions about the appropriateness and utility of analyzing observational data as part of a future drug safety surveillance process.

The proposed project was designed to add to the literature in several important ways, with potential methodological, policy, and clinical implications. First, from a methodological perspective, the study detailed and provided empirical evidence to inform the potential use of a novel method for identifying drug safety issues in automated healthcare databases as part of an active surveillance system. This method leverages advances in pharmacoepidemiology, biomedical informatics, and pharmaceutical sciences to provide an analytical framework that could support continued drug outcome research beyond the scope of this study's ACE inhibitor analyses.

Second, from a policy perspective, the evaluation of how to interpret findings across a network of data sources may have broader implications for initiating the national active surveillance system. There is little research to inform how decision-making processes will accommodate information when generating, strengthening and confirming hypotheses about potential drug-related effects[23]. The role of exploratory analyses in an active surveillance system and the relative confidence in information that can be gained from such analyses is undetermined. Studying heterogeneity across sources and the potential use of a meta-analytic framework to integrate estimates provided insights to inform the governance of the future national active surveillance system about what level of evidence is necessary to take appropriate action about emerging safety issues.

Finally, from the clinical perspective, the exploratory analyses of ACE inhibitors have the potential to generate hypotheses that could shape future understanding about the effects of these medicines. Products that showed comparable safety profiles may stimulate interest in exploring the current inconsistencies in product labeling across the class. Alternatively, products observed to have differential effects, in which case further studies may be warranted to confirm and communicate these differences to inform clinical practice.

# CHAPTER TWO: BACKGROUND AND SIGNIFICANCE

## 2.1 The Increasing Importance of Drug Safety in the Quality of Healthcare

Drug safety continues to be a major public health concern in the United States. In order for patients and health care providers to make effective therapeutic decisions, they need to be informed of the potential benefits and harms of alternative treatment options. While the efficacy of prescription medicines is generally well-characterized from the series of randomized clinical trials conducted during drug development, the safety profile of medicines is often less certain and poorer understood. Research suggests that drug safety information is the highest information priority for patients, and that the perception of side effects is influential in many patients' decisions about taking a medicine[1]. This patient focus is well-justified. Lazarou et al estimated that, in 1994, between 76,000 and 137,000 hospital patients died from an adverse drug reaction (ADR), ranking adverse drug reactions as the fourth to sixth leading cause of death[2], and resulting in health care costs estimated between $3.6 billion[3] and $8 billion[24] annually. The Institute of Medicine report *To Error Is Human*, which claimed 44,000 to 98,000 Americans die each year due to medical errors[4], though not all drug adverse reactions are medical errors, nor do all medical errors result in adverse drug reactions.

In order to understand fully understand the magnitude of the effect of drug safety, it is important to provide the proper context around the potential quality issues associated within this domain. The FDA Task Force on Risk Management provided a framework to classify the sources of risk from medicinal products, shown in Figure 1[6].

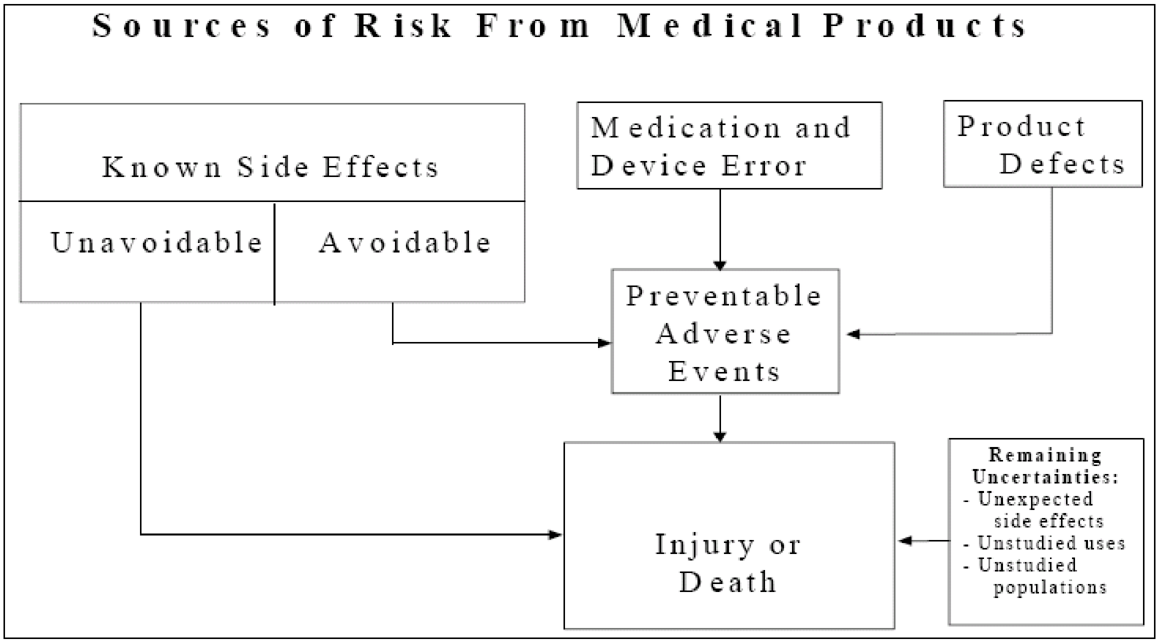


**Figure 1: Sources of risk for medical products**

Most injuries and deaths associated with drug use result from their known side effects. The 'known unavoidable side effects' are typically not regarded as medical errors, but are simply the unfortunate potential consequence of choosing a pharmaceutical intervention in hopes of achieving the benefits of that treatment. While some side effects are unavoidable, more than half of the side effects from pharmaceuticals can be prevented or minimized by careful product choice and use[6]. Two other sources of preventable adverse events include medication/device errors and product defects. Medication or device errors may involve the incorrect administration of the prescribed product or incorrect operation or placement of a medical device[6]. Product defects may be the result of inadequate product quality control and quality assurance during manufacturing. Failure to prevent avoidable adverse effects could certainly be characterized as a medical error. The final category of potential risk characterized in Figure 1 is 'Remaining Uncertainties'. These include unexpected side effects, unstudied uses, and unstudied populations.

Unexpected adverse events are those drug associations not identified prior to regulatory approval, either due to the rare occurrence of the event or the unstudied use of the drug within specific populations. Physicians and patients expect that when medications are prescribed correctly for labeled indications and are used as directed, these medications generally will have beneficial effects and will not cause significant harm. This confidence in pharmaceutical products reflects trust in the effectiveness and integrity of the drug approval and monitoring process[25]. Yet, a 2004 Harris poll showed a sharp decline in public confidence in the FDA, with negative ratings of 58% compared with 37% two years prior[5]. The information around known benefits and risks form the basis of therapeutic decision-making, but unexpected adverse events cannot easily enter into the benefit-risk equation; patients and physicians don't know what they don't know, but they expect the regulatory authorities and manufacturers to tell them what they should know. Different decisions, potentially resulting in improved outcomes, may be made if new information were to be introduced.

The frequency of new information being brought to light is quite striking. A study by the US General Accounting Office (GAO) concluded that 51% of all approved drugs had at least 1 serious adverse drug reaction that was not recognized during the approval process[5]. A revised estimate from 1994-1997 showed 30% of products required significant label changes following introduction[6]. Once the information about a potential safety concern is known and understood, patients and providers can make informed therapeutic decisions. However, the interlude between drug introduction and new safety information being available presents a potential quality concern, as patients may make decisions that incur unnecessary harms that they would otherwise not had they been provided with better quality information. The degree of exposure to drugs during this time of imperfect information may be extensive. Nearly 20 million patients in the United States took at least 1 of the 5 drugs withdrawn from the market between September 1997 and September 1998. Seven drugs approved since 1993 and subsequently withdrawn from the market have been reported as possibly contributing to 1002 deaths[7]. It could be speculated that at least some of these negative outcomes could have been averted had the full safety profile been understood at the time of therapeutic decision-making.

While "the contribution of serious adverse events resulting from unexpected side effects to the overall rate of serious adverse events is relatively small"[6], the level of media attention and public scrutiny of unexpected adverse events is quite significant. Recent notable product withdrawals, such as rofecoxib[26-29], tegaserod[30], and pemoline[31], and other emerging potential safety concerns, such as rosiglitazone[32-35], have increased visibility and skepticism of the effectiveness of the current post-approval safety surveillance processes. With this sensitivity comes the concern that regulatory decision makers may become too conservative in their assessment of the benefit-risk balance of medicines, putting too much emphasis on rare but serious adverse events without sufficient perspective placed on the efficacy profile of the medicine for the indicated population. In this regard, inadequate understanding of the safety of medicines can result in misuse, overuse, and underuse of pharmaceuticals as therapeutic alternatives.

## 2.2 History of FDA's response to drug safety

While unexpected drug adverse events has gained significant attention over the past five years, the issue has been at the forefront of FDA's activities for over 100 years. When the US Federal Food and Drugs Act of 1906 was passed, the primary focus was ensuring drugs were pure and free from contamination, with no requirement of efficacy[36]. Nonetheless, there were 107 deaths in 1937 from the use of diethylene glycol as a solvent for sulfanilamide. Although the toxicity of diethylene glycol was known at the time, it was not known to the manufacturer, and an amendment to the original act was passed in 1938 to outlaw misbranding of ingredients or false advertising claims[36]. The most significant drug safety event occurred in 1961 when published reports identified an association between thalidomide and a 20% increase in fetal malformation and phocomelia[36]. The number of children born with serious congenital malformations as a result of maternal use of thalidomide was estimated between 6,000 and 12,000, with the majority being born in Germany[37]. In 1962, Congress responded by passing the Kefauver-Harris amendment to the US Federal Food and Drugs Act, requiring pre-marketing submission of both efficacy and safety data to the FDA[36]. The FDA also started a systematic collection of reports on all types of adverse drug reactions, originally chiefly through the Hospital Reporting Program[37]. The spontaneous adverse event reporting system is a tradition that continues to this day through the FDA's MEDWatch program, with all case reports archived within the Adverse Event Reporting System (AERS)[38]. Originally, a spontaneous reporting system for suspected adverse effects of drugs was the only conceivable early warning system for possible drug adverse reactions[39]. Since that time, other potential models for adverse event reporting, such as prescription event monitoring, have been proposed and implemented in other countries[39, 40].

Various stakeholders have recognized the need to improve current pharmacovigilance practice and the opportunities that exist to expand the use of observational data in that pursuit[1, 5, 26, 41, 42]. The Institute of Medicine study of the drug safety system was largely prompted from the market withdrawals of troglitazone, cerivastatin, and rofecoxib[43]. In 2007, Congress passed the FDA Amendment Act, which in part, mandated the "establishment of a postmarket risk identification and analysis system" that leverages observational healthcare data, including administrative claims and electronic health records, to monitor approved medicines on a periodic basis[11]. In response, FDA established the Sentinel Initiative, an effort to create and implement a national, integrated, electronic system for monitoring medical product safety[44]. In their initial work, FDA has called for additional research to inform the "science of safety" and establish best practices for the appropriate use of observational data and analysis within an active surveillance system.

## 2.3 Approaches for evaluating drug safety issues

Prior to regulatory approval while a drug is in development, one of the primary sources of safety information about medical products is clinical trials. Randomized experiments are designed and conducted to test the efficacy of the drug, typically in comparison to placebo or standard of care. During the course of these efficacy trials, adverse events are captured at each study visit, and final study reports typically summarize these events as frequency tables. Typically, observation of serious adverse events during clinical development is cause for study termination unless the benefits can be shown to outweigh the potential risks. Other adverse events captured during development commonly reflect nuisance side effects, such as headache and nausea, for which a causal relationship may be undetermined but general consensus suggests any purported relationship would not alter therapeutic decision-making. Randomized experiments are generally regarded as the highest level of evidence, as studies lead to an unbiased estimate of the average treatment effect[45].

Unfortunately, most trials suffer from insufficient sample size and lack of external validity to reliably estimate the risk of other potential safety concerns for the target population[1, 41]. Rare side effects and long-term outcomes (both positive and negative) may not be known when a product is approved because of the relatively small size and short duration of clinical trials. For products intended to treat chronic, non-life-threatening conditions that occur in large populations, the International Conference for Harmonization (ICH) recommends a baseline safety database that typically involves at least 1,500 patients with at least 6 month exposure time to reliably (95 percent of the time) identify events happening at the 1-percent level[6]. In other words, events that occur less frequently than 1 in 100 patients are not expected to be detected under this recommendation. Adverse events that occur in specific populations (like children, pregnant women, elderly, and patients with other comorbid conditions) may not be detected in clinical trials because these subgroups are not studied as comprehensively in drug development[23]. For a clinical trial to provide the appropriate insights for a particular safety question, the choice of outcomes, the duration of treatment, length of follow-up, target population, and statistical power, must all be correctly specified. Due to these limitations, it is generally accepted that safety can only be regarded as provisionally established at the time of approval and knowledge about the safety profile will continue to be developed in clinical practice[23].

Meta-analysis of clinical trial data has gained favor as one approach for overcoming the limitation of insufficient sample in any one study. Adverse event rates for both the treated and untreated study arms can be derived from samples, and composite estimates of relative effects can be produced by weighting studies by the inverse of the variance of each study-specific measure[46]. Numerous methods exist for meta-analysis to pool study-level effects[47], and are commonly applied in systematic reviews that assess medical product efficacy and effectiveness. Meta-analyses have the advantage of increasing power and improving precision, and offer the ability to answer questions not posed by individual studies or explore conflicting claims generating by different experiments [47]. Meta-analyses have been applied in drug safety contexts to generate estimates for specific events once concerns came to light. Pooled analysis of rofecoxib data has been shown retrospectively to detect a significant safety signal with acute myocardial infarction three years prior to the product withdrawal[48]. Nissen and Wolski conducted a meta-analysis of rosiglitazone clinical trials to identify potential increase in cardiovascular events[33]. One challenge impeding its broader use as an exploratory tool is that proper meta-analysis requires careful consideration of specific study designs, within-study biases, variation across studies, and reporting biases that may be present when interpreting analysis results. It has been noted also that "the interpretation of the results of systematic reviews with meta-analyses includes a subjective component that can lead to discordant conclusions that are independent of the methodology used to obtain or analyze the data"[49].

While large-scale clinical trials and pooled meta-analysis results are often desirable to produce the most reliable measure of an effect, they are often infeasible logistically or ethically. Observational studies provide an alternative approach to evaluating drug safety questions that can provide the necessary information about the drug effects to support clinical decision making. Depending on the questions posed, a primary analysis of an appropriate observational study may provide better information than the analysis of an existing clinical trial data set.[50] Observational studies provide empiric investigations of exposures and the effects they cause, but differ from experiments in that the there is no control of assignment of treatment to subjects[45]. Observational studies can take many forms of epidemiologic investigation, using different methods for data collection, applying alternative study designs, and leveraging different analysis strategies[51-56]. These studies can range from population-based cohort studies with prospective data collection to targeted disease registries to retrospective case-control studies.

One type of resource that has provided fertile ground for epidemiologic investigation has been observational healthcare databases. Administrative claims and electronic health record databases have been actively used in pharmacoepidemiology for over 30 years[57], but have seen increased use in the past decade due to increased availability at lower costs and technological advances that made computational processing on large-scale data more feasible. Observational healthcare databases offer researchers the opportunity for secondary use of data captured as part of the healthcare delivery system to study effects amongst any observed medical products. Many such databases contain large numbers of patients that make it possible to examine rare events and specific subpopulations that previously could not be studied with sufficient power[58]. The large population size make it possible to estimate absolute incidence rates across a wide array of potential outcomes and to measure amount of exposure in a large population to produce more accurate measures of potential public health impact[59]. Because the data reflects healthcare activity within a real-world population, it offers the potential to complement clinical trial results which suffer from lack of generalizability. Long-term longitudinal capture of data in these sources can enable studies to monitor the performance of risk management programs over time[60].

Administrative claims databases have been the most actively used observational healthcare data source. Administrative claims databases typically capture data elements used within the reimbursement process. Providers of health care services such as physicians, pharmacies, hospitals, and laboratories submit encounter information so that they will be paid for these services[61]. This commonly includes pharmacy claims for prescription drug fills (providing what drug was dispensed, the dispensing date, and the days supply), and medical (inpatient and outpatient) claims that detail the date and type of service rendered. Medical claims typically contain diagnosis codes used to justify reimbursement for the procedures. Age and gender can also commonly be inferred from the available data. In these databases, data are recorded only when a patient has a reimbursable encounter with the health care system that has been properly filed, coded and adjudicated by the payer.[14] As a result, many key data elements may not be available. Information on over-the-counter drug use and in-hospital medication is usually unavailable and the patient's compliance with the prescription is generally unknown[62]. Retail pharmacy claims data can be used to study drug utilization pattern, but the completeness of these data can vary by patient age[63] or other unobservable characteristics. Claims can be aggregated by payers, healthcare systems, or data aggregators, though each may have a different perspective on how to define observation periods (whether it be the time insured, the time in the system, or simply the span of time that data was observed). While the databases over longitudinal coverage, the amount of times that patients persist within a given database can vary significantly. This problem can be especially pronounced in payer databases; for example, it is estimated that health maintenance organizations (HMO) have an annual turnover rate of 20% to 30%[64]. Therefore, a database capturing healthcare encounters may contain records spanning a decade or more, but the average person may only exist in the database for 18 to 24 months.

Electronic health records (EHR) generally contain data captured at the point of care, with the intention of supporting the clinical process. A patient chart may include demographics (birth date, gender, and race), height and weight, and family and medical history. Many EHR systems support provider entry of diagnoses, signs, and symptoms, and also capture of other clinical observations, such as vital signs, laboratory values, and imaging reports. Beyond this, electronic medical records may often contain findings of physical examinations and the results of diagnostic tests[14]. EHR systems usually also have the capability to record other important health status indications, such as alcohol use and smoking status[65], but the data may be missing in many patient charts[61]. Unless integrated across an entire health system, electronic health record systems are generally maintained independently by physician practices. The provider and office staff enter information elicited from the patient or generated by the physician, but are also responsible for entering relevant clinical information from services rendered outside the practice, including conditions diagnosed by outpatient specialist physicians or during hospital admissions[61]. Drug exposure may be inferred from various sources; providers may use the EHR system to capture patient-reported medication history and/or to write prescriptions, but there may be no confirmation that prescription was filled at a pharmacy. As a result of discontinuous care within the US health care system, a patient may have multiple electronic health records scattered throughout the providers they've seen, but rarely are those records integrated together, so each reflect a different and incomplete perspective of that person's healthcare experience.

For both administrative claims and electronic health records, drug safety analyses are considered a secondary use of the data. Therefore, the onus is on the researcher to fully understand and assess the relative strengths and limitations of each potential source, prior to conducting an evaluation. Data recorded in either system reflects data used for its primary intent and therefore may not necessarily represent the information desired for study. For example, diagnoses recorded on medical claims are used to support justification for the payment of a given procedure; this diagnosis could represent the condition that the procedure was used to 'rule out' or can be an administrative artifact of being the code used by a medical assistant to maximize the reimbursement amount. Similarly, patients without a diagnosis recorded do not necessarily reflect the absence of a condition, as the code may not be used due to lack of seriousness or convenience to facilitate payment procedures. A similar limitation exists in EHR systems, where in addition to concerns about incomplete capture, data may be artificially manipulated to serve clinical care. For example, physicians may neglect to remove conditions that have subsided, or may remove many records all at once to make viewing the problem list in the electronic system more convenient. Some diagnosis codes have been studied through source record verification and have demonstrated adequate performance characteristics[66-76], with other conditions and systems are less certain[77-80]. Most systems have insufficient processes to evaluate data quality a priori, requiring intensive work on behalf the researcher to prepare the data prior to analyis[81]. Both types of sources require inferences to estimate potential drug exposure. Inferences can be made in administrative claims sources based on pharmacy dispensing records, while inferences for HER systems rely on patient self-report and physician prescribing orders[61]. Neither reflect the timing, dose, or duration of drug ingested, so assumptions are required in interpretation of all study results.

The principle concern for all observational studies, which is of particular relevance in observational database evaluation, is the potential for bias. Schneeweiss et al illustrated some of the potential sources of bias that are introduced throughout the data capture process for both administrative claims and electronic health records, as shown in Figure 2[14].

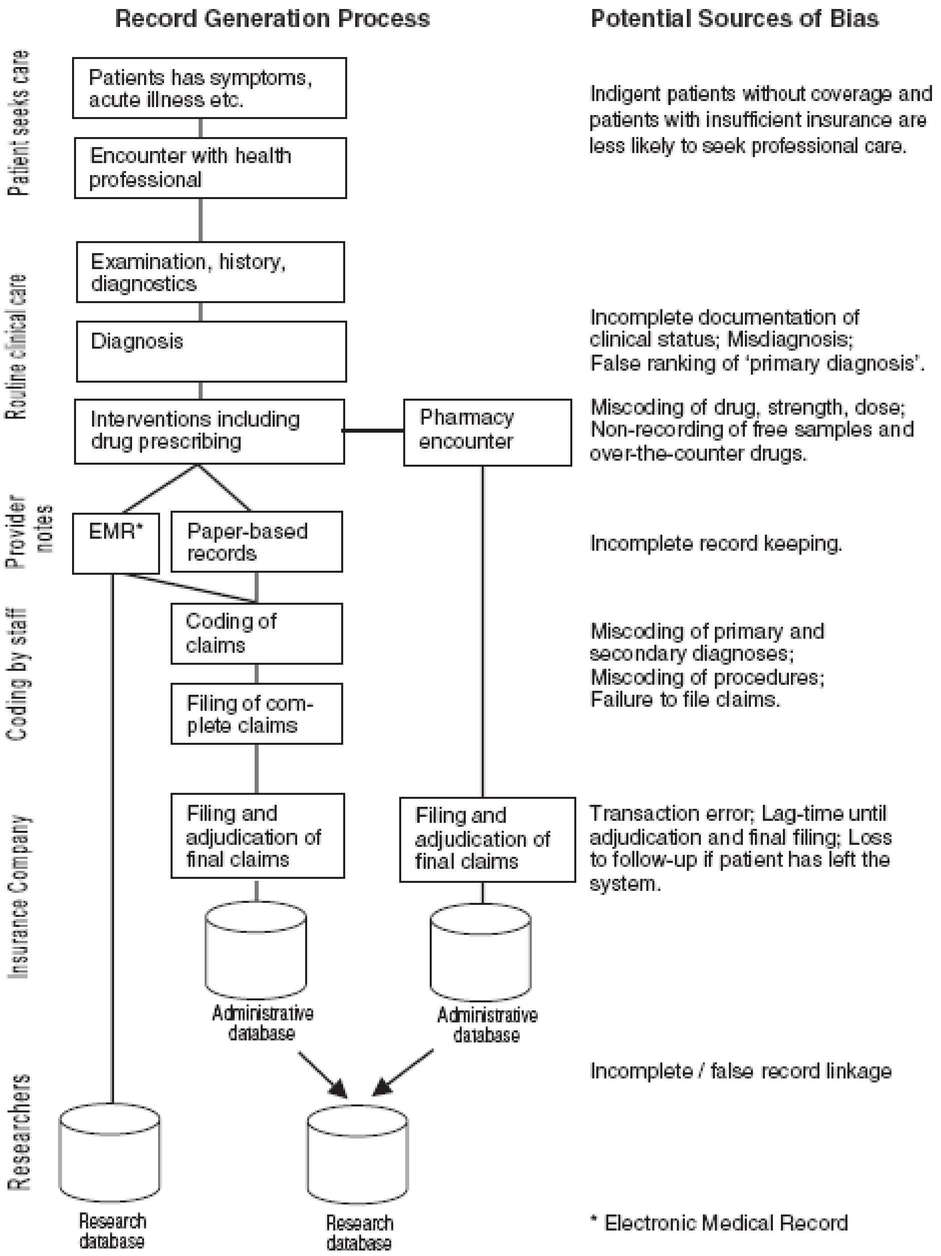

Record Generation Process
Potential Sources of Bias
Patient seeks care
Routine clinical care
Provider notes
Coding by staff
Insurance Company
Researchers
Patients has symptoms, acute illness etc.
Encounter with health professional
Examination, history, diagnostics
Diagnosis
Interventions including drug prescribing
Pharmacy encounter
EMR*
Paper-based records
Coding of claims
Filing of com- plete claims
Filing and adjudication of final claims
Filing and adjudication of final claims
Administrative database
Administrative database
Research database
Research database
Indigent patients without coverage and patients with insufficient insurance are less likely to seek professional care.
Incomplete documentation of clinical status; Misdiagnosis; False ranking of 'primary diagnosis'.
Miscoding of drug, strength, dose; Non-recording of free samples and over-the-counter drugs.
Incomplete record keeping.
Miscoding of primary and secondary diagnoses; Miscoding of procedures; Failure to file claims.
Transaction error; Lag-time until adjudication and final filing; Loss to follow-up if patient has left the system.
Incomplete / false record linkage
* Electronic Medical Record

**Figure 2: Generation of health care utilization databases and potential sources of errors and bias[14]**

An observational study is biased if the treated and control groups differ prior to treatment in ways that can influence the outcome under study[45]. Several forms of bias can arise through a study. In the context of drug safety analyses, one of the most challenging issues of confounding by indication: a medical product is differently used as treatment for a given disease, but factors associated with the underlying disease independently influence the risk of outcome[82]. Therefore, a medical product can appear associated with the outcome without appropriate control for the underlying condition, and confounding may persist even despite advanced methods for adjustment[83]. Confounding can also exist due to predisposition for healthcare utilization, either due to functional status, or access due to proximity, economic and institutional factors[84]. An additional concern is immortal time bias, whereby outcomes are not observable within the defined time-at-risk[85-87].

Several strategies exist for minimizing the effects of bias within observational database studies. These include design-level considerations and analysis approaches. One design strategy is to impose restrictions on the sample selected to increase validity, potentially at the expense of precision. These restrictions are quite analogous to clinical trials, and include ensuring incident drug use, similar comparison groups, patients without contraindication, and comparable adherence, as shown in Figure 3 [88]. Schneeweiss et al showed in an example of statin and 1-year mortality how bias was minimized at each stage of restriction. The restriction to incident users deserves special attention. Use of a new user design can minimize prevalent user bias and eliminate selection of intermediate variables[89-91]. Within a new user design framework, measures of effect focus on events occurring after the first initiation of treatment, which allows a more direct comparison to a comparator group using an alternative treatment. The design can be logically extended to study drug switching and add-on therapies, as long as incident use of the target drug is preserved[90].

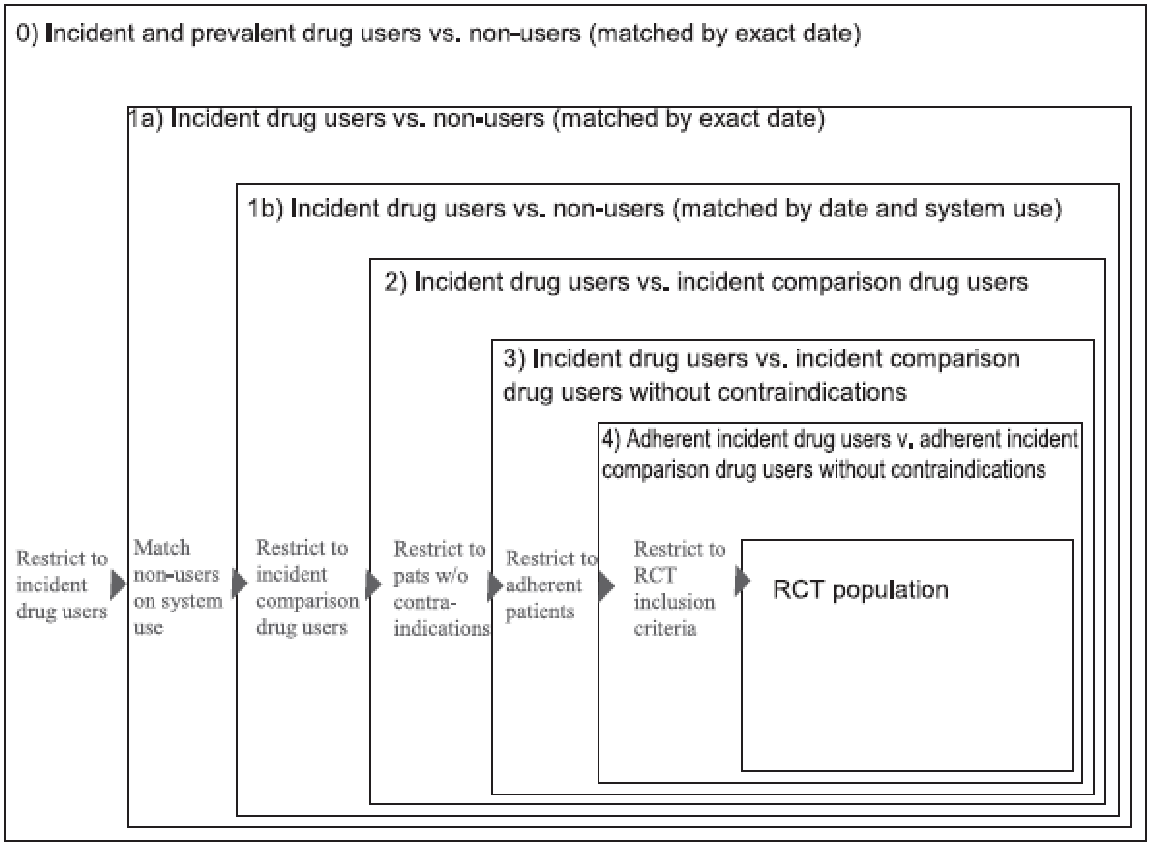

0) Incident and prevalent drug users vs. non-users (matched by exact date)
1a) Incident drug users vs. non-users (matched by exact date)
1b) Incident drug users vs. non-users (matched by date and system use)
2) Incident drug users vs. incident comparison drug users
3) Incident drug users vs. incident comparison drug users without contraindications
4) Adherent incident drug users v. adherent incident comparison drug users without contraindications
Restrict to incident drug users
Match non-users on system use
Restrict to incident comparison drug users
Restrict to pats w/o contra-indications
Restrict to adherent patients
Restrict to RCT inclusion criteria
RCT population

**Figure 3: Population restrictions to approximate clinical trials in observational studies**[88]

Comparator selection is also an important design consideration to reduce confounding by indication. The comparator definition should yield patients in the same health circumstance as those eligible to be new users of target medication. In some regards, when assessing a drug safety issue, the comparator is desired to represent the 'standard of care' that would be provided to that patient had they not been prescribed the target drug, such that relative effect estimates represent risk above and beyond that that patient could otherwise expect. A challenge in comparator selection comes when there is no truly comparable standard of care to evaluate against, or when there is significant channeling bias influencing treatment decision to a particular drug class. In this regard, evaluation studies can be highly sensitive to the comparator selected, and a criticism of these studies is often the subjective nature by which the comparator was selected.

Once a design is established, bias can be further minimized through analysis strategies, such as matching, stratification, and statistical adjustment. Variables commonly considered for adjustment are those which are observed to have different baseline characteristics, or are known to have the potential to influence treatment decisions or outcome occurrence. These may include patient demographics, such as age, gender, and race. It may also include patient comorbidities, either expressed as a set of binary classifiers of specific diseases or as a composite index of comorbidity. One commonly used measure is the Charlson index[92-102], which was originally developed to predict mortality, but has also been shown to be related to healthcare expenditures[103]. Adjustment for comorbidity index has shown to be useful for exploratory data analysis[104], but are not sufficient to address all potential sources of confounding due to background conditions. Additional variables often cited include prior use of medications, and markers for health service utilization, such as number of outpatient visits and inpatient stays. The specific definition and application of these covariates is highly variable across drug safety evaluation studies. It has been shown that covariate selection can influence effect measures, regardless of the modeling approach undertaken, particularly if effect modification exists[105].

Once variables are identified, they can be controlled for through direct matching or stratification, whereby the target and comparator groups are logically divided by the attributes of the covariates. However, in a multivariate context, the data may be too sparse to provide adequate sample to match on all covariates or provide subpopulations within each strata. A popular tool to overcome this limitation is propensity score analysis[45, 106].

Propensity scores are most commonly used in cohort studies. Within the context of cohort studies, the propensity score is estimated as the conditional probability of treatment assignment, given the observed characteristics prior to exposure. The propensity score provides a scalar value that summarizes all covariates, commonly estimated through logistic regression, which can then be used for matching or stratification[107]. Propensity scores can be been shown to balance the distribution of covariates between two cohorts, although patient-level covariate values may differ within paired groups[45]. Variables introduced in the propensity score model which are confounders (related to both exposure and outcome) or related to outcome alone have shown to minimize bias in outcome effect measures[108]. While propensity score adjustment has increased in popularity, the practical effect, in relation to typical multivariate modeling approach, can be modest in many circumstances[109]. However, its use has several desirable characteristics that make its choice preferred to conventional approaches, including the focus on pre-exposure characteristics, improved balance, and better control of confounding that could influence rare outcomes or small relative effects[110].

As with other approaches, the propensity score model is only as good as the covariates selected to provide the adjustment. Propensity score may balance observed confounders, but does not balance in factors not incorporated into the model. This is a particular problem for analysis of electronic healthcare databases, where many important covariates, such as smoking status, alcohol consumption, body mass index, and lifestyle and cultural attitudes to health, are not captured. Sturmer et al demonstrated that further adjustment could be achieved by conducting supplemental validation studies to collect additional information on previously unmeasured confounders[111]. Schneeweiss showed how unmeasured confounders biased estimates of COX-2 inhibitors and myocardial infarction[112].

Seeger et al highlighted how a model without the appropriate variables included could yield a biased estimate in a case study exploring association of statin therapy and myocardial infarction[113-115]. Strategies for automated selection of large sets of covariates have been proposed as potential solutions to minimize risk of missing an empiric confounder[116]. Sensitivity analysis has been proposed as an additional approach to assess the potential consequences of unobserved confounding[117], but is unfortunately rarely reported in published studies.

Instrumental variable (IV) analysis presents a potential solution to adjusting for uncontrolled confounding through control of a factor that is related to exposure but unrelated to outcome[118-120]. Several studies have shown how instrumental variable analysis can reduce bias[121-124]. A challenge in IV analysis is identifying a covariate that satisfies the criteria of an instrumental variable, particularly with regard to having no effect on the outcome. For active surveillance, where multiple outcomes may be explored for a given outcome, the selection of a common instrumental variable becomes even harder.

One consideration for all statistical adjustment techniques in drug safety evaluation studies is the danger of introducing bias. Statistical control for variables which either increase bias or decrease precision without affecting bias can produce less reliable effect estimates[125]. For example, bias can also be induced if an analysis improperly stratifies on a collider variable[126]. As a result, care has to be taken in any evaluation study to develop a parsimonious model which maximizes the bias control while minimizing the risk of introducing bias or inflating variance.

In pharmacoepidemiology circles, the strategies for overcoming the limitations in studying a given observational healthcare database have been well understood, and best practices are gaining widespread agreement. One emerging area of opportunity for drug safety evaluation involves applying these same practices across networks of observational databases. The potential value of a network-based approach is appealing. While one database can be large in size, study restrictions can result in insufficient sample of the population of interest to provide reliable estimates of drug-related effects. Conducting analyses across multiple sources can alleviate concerns of insufficient sample size, and also provide higher quality evidence by allowing effects to be evaluated concurrently within disparate source populations. Several efforts have shown promise in constructing networks of databases and conducting evaluation studies across the network. The HMO Research Network is the most notable example. The network was established in 1994 and is comprised of 16 HMO organizations covering over 15 million persons, with each organization maintaining administrative claims data that can be pooled across the network for specific evaluation studies. Researchers within the network have conducted various studies to support public health, including drug safety evaluations[127-135]. Meningococcal Vaccine Study used a network of administrative claims databases to conduct a cohort study of Guillian Barre syndrome following meningococcal vaccination[136]. Some preliminary work has shown how analyses can be successfully executed across such a network[137-139]. In particular, Rassen et al demonstrated how propensity score techniques could be applied in a distributed setting to provide adjusted effects while minimizing concerns of patient privacy[139]. These prior successes have led to active discussions about how to establish a distributed network that could focus on drug safety evaluations as part of a national active surveillance system[140-143].

## 2.4 Approaches for identifying potential drug safety issues

FDA Guidance for Industry: Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment defines pharmacovigilance as "all scientific and data gathering activities relating to the detection, assessment, and understanding of adverse events. These activities are undertaken with the goal of identifying adverse events and understanding, to the extent possible, their nature, frequency, and potential risk factors.[144]" The principle concern of pharmacovigilance is the timely discovery of adverse drug reactions that are novel in terms of their clinical nature, severity and/or frequency as early as possible after marketing, with minimum patient exposure[145]. The ultimate goal of pharmacovigilance is the rational and safe use of medicines; the findings are intended to influence physicians, pharmacists, and patients in their choice of medicines (including self-medication) and the precautions to be taken[39].

Methods and processes for evaluating specific drug safety issues once identified have been well established and expanded use of observational studies continues to be refined, as described in the previous section. Once a signal is detected, a thorough and manually-intensive evaluation is conducted, requiring use of many information sources, including pre-clinical studies, clinical trials, spontaneous adverse reaction reports, epidemiological studies, and data collected for other purposes[146]. The course of events leading to the identification and evaluation of adverse events "frequently follows an S-shaped curve, with 3 major phases: a latent period during which a suspicion arises at some point, followed by the often sudden accumulation of data (signal strengthening) and, finally, a usually lengthy phase of evaluation during which the adverse effect is confirmed (signal testing), explained and quantified"[39]. Following signal evaluation, if it is deemed that there is reasonable evidence to suggest an association between a drug and an adverse event, a decision about the effect on overall safety must be made and the appropriate actions taken, including 1) to change a product label, 2) to conduct patient or physician education, 3) to limit advertising to patients or physicians, 4) to modify approved indications, 5) to restrict use to selected patients, 6) to conduct additional post-marketing studies or trials, and 7) to suspend marketing or immediately withdraw a drug[147]. The FDA recognizes the importance of timely communication of emerging safety information, stating that "informing healthcare providers of changes and updates in information about pharmaceuticals during the post-marketing period is essential to assuring continued safe use of these drugs. It is critical that physicians understand and act on the latest information available regarding the appropriate use of a medication[148]." Because signal detection is the first step on the cascade of steps that lead to communications aimed at improving patient-provider decision-making, it could be expected that enhancing the ability to identify potential safety concerns of medicines can result in downstream improvements in patient quality of care. However, one of the largest gaps in our current system is how to identify the issues that warrant the evaluation.

Once a drug has been approved and is introduced on the market, the FDA's "postmarketing surveillance programs focus primarily on (1) identifying events that were not observed or recognized before approval, and (2) identifying adverse events that might be happening because a product is not being used as anticipated"[5]. Spontaneous adverse event reporting remains the cornerstone of pharmacovigilance activities. The FDA receives about 400,000 reports annually, primarily from drug manufacturers who are required to report serious, unexpected safety events within 15 days, and a minor proportion coming directly from health care providers and patients[7]. The chief use of spontaneous adverse event reports is to facilitate clinical review of case series. A case series of related events is often the first initial warning that a potential association may exist, and a single reliable case report may be sufficient to provide definitive evidence about a rare, serious idiosyncratic event[149]. While case series are generally regarded as low on the hierarchy of evidence, behind observational cohort studies and randomized trials[150], the use of case series analysis has a prominent role when other information cannot be made available[151]. At the same time, clinical review constitutes the primary bottleneck: with hundreds of thousands of reports submitted to FDA, WHO and other organizations each year, every one of which cannot possibly be reviewed by the available experts[152].

Although the careful review of pharmacovigilance experts remains central to the drug safety process, statistical data mining algorithms are becoming increasingly popular supplementary tools for safety reviewers[24, 25]. Various disproportionality analysis methods exist, but each approach attempts to answer the same question: which drug-event combinations are reported more frequently than we would have expected if the drug and event were truly independent[153]? The proportional reporting ratio (PRR) was first proposed by Evans et al as a simple tool to help prioritize amongst the potential relationship identified within a spontaneous adverse event reporting databases [154]. The Reporting Odds Ratio (ROR) was also established[155], and is used in some countries in Europe, including the Netherlands[156, 157]. Multi-item Gamma Poisson Shrinker (MGPS) was conceived as a Bayesian approach to the disproportionality problem and also incorporated stratification by age, gender, and year-of-report in an attempt to both shrink small estimates and minimize potential sources of bias when calculating an expected value to compare to[9]. MGPS has become the preferred method of choice at FDA, and actively used throughout the pharmaceutical industry[158-160]. Bayesian confidence propagation neural network approach, and its Information Component (IC) metric, was developed as another Bayesian approach and is currently used at WHO[161-167]. While most agree the potential for bias is significant, preliminary work using IC shows stratification by age and gender may decrease performance of spontaneous data mining[168]. Additional work has looked to expand into using these methods for drug-drug interactions[169, 170].

Data mining has shown that it does not detect safety issues sooner than a simple heuristic of 3 reported events, but the proportion of true relationships is higher[171]. After the methods had widespread use and different approaches were gaining favor in different corners of the world, a body of research was conducted to compare the performance for the approach. Preliminary work showed the methods had comparable performance[153, 172]. Some studies observed that performance differences between the PRR and MGPS methods are related to stratification effects, tradeoffs in sensitivity and specificity, and inequities in the thresholds that have been adapted for each method. PRR was shown to be more sensitive and less specific than MGPS[173]. With all approaches however, false positives present the most significant challenge. Hochberg et al showed “there is a substantial fraction of SDRs for which no external supporting evidence can be found, even when a highly inclusive search for such evidence is conducted”[174]. It is generally understood that the weaknesses of spontaneous adverse event reporting cannot be overcome by data mining methodologies alone[8, 145, 146, 153, 175-182].

Experience gained internationally shows that spontaneous reporting is effective in providing information about a wide range of different adverse effects and other drug-related problems. It has been mainly helpful in detecting adverse events that are often allergic or idiosyncratic reactions, characteristically occurring in only a minority of patients and usually unrelated to dosage and that are serious, unexpected and unpredictable, and unusual effects that are related to the pharmacological effects of the drug and are dosage-related[183]. On the other hand, spontaneous reporting is of less use for the study of adverse effects with a relatively high background frequency and occurring without a suggestive temporal relationship[39].

While the spontaneous adverse event reporting system has value in generating hypotheses about potential associations, it  has several limitations that make causal assessments difficult: voluntary reporting suffers from chronic underreporting and maturation bias, and the unknown nature of underlying population make true reporting rates difficult to obtain and use for comparisons. It has been estimated that only about 1% of all adverse drug reactions and about 10% of all serious adverse drug reactions are reported[5]. Reports are "usually based on suspicion, and may be preliminary, ambiguous, doubtful or wrong"[39].

Recognizing the limitations of spontaneous adverse event reporting, various efforts have sought to leverage observational healthcare databases for event detection. The CDC has played a leading role in establishing public health surveillance programs to inform medical product safety issues. The National Electronic Injury Surveillance System-Cooperative Adverse Drug Event Surveillance System has enabled monitoring of adverse drug events leading to emergency department visits[184-186]. Another successful project has been the CDC Vaccine Safety Datalink (VSD), which demonstrated the feasibility of establishing a distributed network of administrative claims sources and conducting systematic analyses to detect vaccine-related adverse events[187]. The sequential probability ratio test was applied to detect increases in intussusception following introduction to the rotavirus vaccines, as well as decreases in several events after the changeover from the whole cell pertussis vaccine to the acellular pertussis vaccine. In these instances, as with spontaneous reporting, the public health surveillance objective is to identify cases of serious events following exposure that wouldn't otherwise be expected. A primary distinction between spontaneous reporting and Vaccine Safety Datalink is that the spontaneous reporting system captures all potential events of any origin, whereas studies designed for VSD focus on a restricted set of specific adverse events known to be potentially caused by vaccines. In that respect, the VSD approach still fall within an evaluation paradigm, whether the system must first be presented within a prior hypothesis of a specific drug-condition relationship and craft an analysis to assess the purported effect. The method was since enhanced[188] and applied to drug safety surveillance as part of the HMO Research Network, which demonstrated the ability to identify known drug-safety issues, including acute myocardial infarction risk following exposure to rofecoxib[137, 138]. However, as the authors note, these studies suffer from several methodological limitations, notably failure to fully address confounding and length of exposure. Also, as with the applications to public health surveillance, these methods have not yet been applied in an exploratory framework to generate hypotheses but are instead applied to targeted drug-condition pairs. As such, 'identification' of the rofecoxib-myocardial infarction effect comes without regard to how many other false positive cases may be detected when using the same approach.

As we move into active drug safety surveillance, the goal shifts from case detection to association detection. That is, the interest in the system expands beyond detecting rare, idiosyncratic events that would not be expected to be seen without exposure, to detecting elevations in risks of conditions that occur in the background population. Where clinical trials may be sufficient for detecting strong associations with highly prevalent outcomes (such as nuisance side effects like headache and nausea), and spontaneous reporting and public health surveillance tools may serve the purposes for identifying cases of rare events (such as Stevens-Johnson syndrome, toxic epidermal necrolyis, and Guillain Barre syndrome), the largest opportunity for an active surveillance system rests in complementing those systems in the gap in between. This may include adverse events that are less commonly observed in clinical trials, that have weaker associations to exposure, and are observed sufficiently often in the general population that case series may not be sufficient to fully characterize the relationship. Several notable adverse events that fall within this category include acute myocardial infarction, fracture, gastrointestinal bleeding, suicidality, and renal and hepatic dysfunction.

While there is a lot of excitement for the potential of an active surveillance system for hypothesis generation, it is widely recognized that significant methodological research is needed to inform the appropriate use of observational data and analysis methods before a national system can be reliably used. Several methods have emerged that attempt to assess multiple outcomes within an exploratory framework across observational databases. A recent FDA-commissioned report summarized a selection of alternative signal detection methods and their potential application to observational data[17], and the Observational Medical Outcomes Partnership also produced a review of methodological considerations for active surveillance that was intended to inform the ongoing scientific dialogue[16].

Curtis et al adapted the empirical Bayes Multi-item Gamma Poisson Shrinkage (MGPS) algorithm to longitudinal administrative claims data, and applied it to the Medicare Current Beneficiary Survey to study effects of COX-2 inhibitors and Non-steroidal antiinflammatory drugs (NSAIDs). The method was used to simultaneously evaluate 259 outcomes and compared to a parallel analysis using traditional epidemiologic methods to assess the concurrent validity of the data mining approach. The authors showed some consistency in cardiovascular effects but also identified several diagnoses that “likely represented indications for the drug”[189].

Noren et al have similarly adapted the Information Component to be applied to longitudinal data, with their Temporal Pattern Discovery method[152, 190]. The method was applied to the UK IMS Disease Analyzer database, which contains electronic health records for 2 million patients through the United Kingdom. Studies successfully demonstrated detection of nifedipine effects of flushing and localized swelling, while providing a visual mechanism to identify potential confounding by indication in effects such as omeprazole and acute pancreatitis[190].

I3 Drug Safety, a subsidiary of United Health Group, developed a commercial tool, i3 Aperio, that is marketed as an active surveillance system[191]. While little of the specific implementation details are publicly available, one article describes the approach as a cohort study that executed against the i3 research claims database[192]. A target drug is compared to a chosen comparator product. Patients are matched with a greedy algorithm based on a propensity score calculated by a logistic regression that includes as covariates age, sex, geographic regions, costs, diagnoses (defined by 3-digit ICD9 codes), drugs, visits, procedures, and labs[192]. Relative risk estimates are provided for 1 year following exposure, with outcomes defined by 4-digit ICD9 codes[192]. The study showed no association between exenatide and acute pancreatitis, though effects of other outcomes were not reported and performance characteristics of the entire system were not provided.

Observational screening is a method originally developed at GlaxoSmithKline and now made commercially available as part of the SAEfetyWorks® software application by ProSanos[18, 193-196]. Screening applies a basic cohort design to estimate the relative rate of condition occurrence among exposed populations compared to the overall population. The method was studied across 1391 labeled events across 10 drugs, and showed 39% sensitivity and 85% specificity when using a threshold requiring two databases to both show a significant effect[18]. SAEfetyWorks introduces two noteworthy innovations: 1) a computationally efficient method for estimating unadjusted incidence rate ratios for all potential outcomes across a wide array of medical products, and 2) a framework for applying methods across disparate data sources to produce composite measures of effects based on threshold criteria imposed across the network of sources.

A common challenge across all methods is determining how to manage the potential false alarms when exploring such a large set of potential outcomes, and determining when evidence is sufficiently compelling to warrant follow-up[23]. To date, no empirical studies have demonstrated the performance characteristics of these methods across a large sample of drug-event pairs, or quantitatively identified the incremental value in supplementing current pharmacovigilance practice with these new methods, either in terms of identifying new issues or faster time-to-detection.

As such, several efforts have begun work to conduct methodological research to develop and study the potential use of analysis methods across an active surveillance system. In 2010, FDA awarded a contract to Harvard Pilgrim Health Care to develop a pilot project, dubbed 'mini-Sentinel', to begin to explore scientific and operational aspects of initiating a national active surveillance system[197]. International efforts are also ongoing to evaluate the potential use of observational healthcare databases for detecting potential drug safety issues. The EU-ADR project “aims to develop an innovative computerized system to detect adverse drug reactions (ADRs), supplementing spontaneous reporting system [by] exploit clinical data from electronic healthcare records (EHRs) of over 30 million patients from several European countries (The Netherlands, Denmark, United Kingdom, and Italy). A variety of text mining, epidemiological and other computational techniques will be used to analyze the EHRs in order to detect ‘signals’ (combinations of drugs and suspected adverse events that warrant further investigation)”[198-201]. The IMI PROTECT (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium) initiative started in 2010 as a collaborative European project to address limitations of current methods in the field of pharmacoepidemiology and pharmacovigilance[202].

The Observational Medical Outcomes Partnership (OMOP) was established in 2009 as a public-private partnership to conduct methodological research to inform the establishment of a national active surveillance system[203]. OMOP is chaired by FDA, managed by the Foundation for the National Institutes of Health, and supported by the pharmaceutical industry, with broad participation from government, academia, payers, healthcare systems, and patient groups, across multiple disciplines, including epidemiology, statistics and medical information, and across the applied health sciences. OMOP consists of a two year research program to evaluate the feasibility and utility of alternative analysis methods and observational health care databases for identifying and evaluating safety and benefit issues of drugs already on the market. OMOP has established a data community of 10 disparate data sources, comprising over 200 million lives in aggregate, and designed a large-scale methodological experiment where a library of methods will be applied to each database and tested against a defined set of test cases (positive and negative controls for 10 medical products) to empirically measure performance in identifying known safety issues and discerning from false positive findings.  All the methods and tools developed by OMOP have been placed in the public domain, and researchers have been encouraged to use these products to advance their own research pursuits.

### 2.5 An integrated active surveillance system within a causal inference framework

While still in its infancy, there is much debate about the intended design and scope of a national active medical product safety surveillance system[204, 205].  An active surveillance system will involve a systematic process for analyzing multiple observational healthcare data sources, including administrative claims and electronic health records, to better understand the effects of medical products by estimating temporal relationships between exposure and outcomes.  The active surveillance system can be used to 1) characterize known side effects, 2) monitor preventable adverse events, and 3) explore remaining uncertainties.  The goal of the active surveillance system is to contribute supplemental information to other existing sources of safety information (including pre-clinical data, clinical trials, and spontaneous adverse event reporting) to support decision-making about appropriate use of medical products for regulatory agencies, providers, and patients.

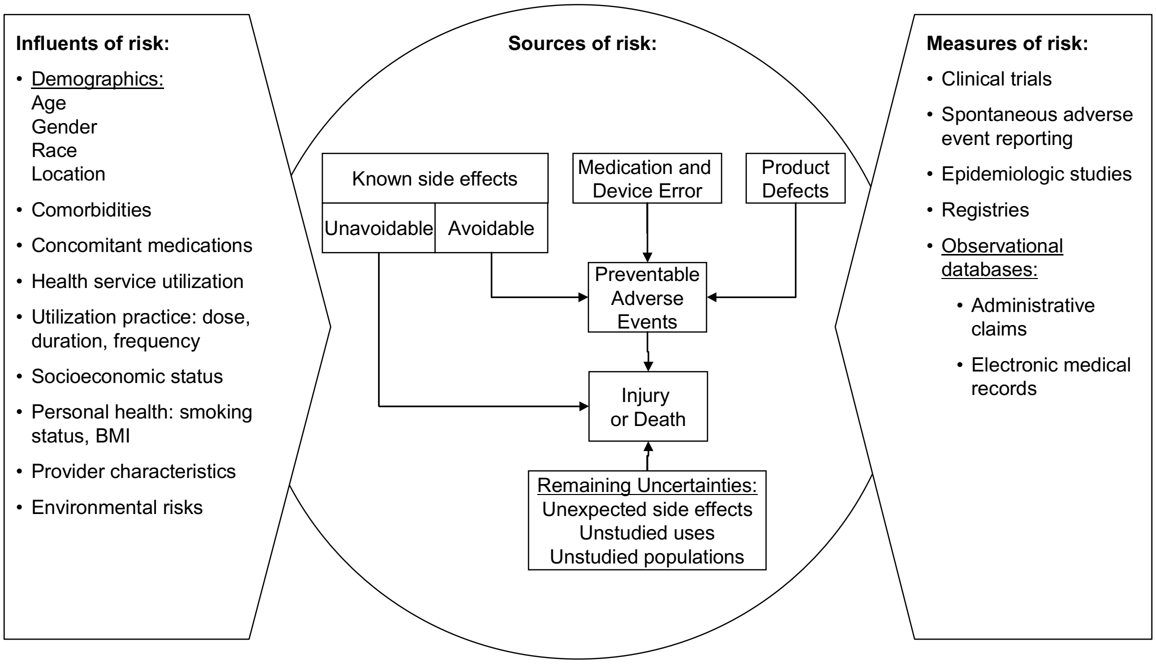


**Figure 4: Conceptual framework for active surveillance**

Figure 4 provides a conceptual framework for active surveillance. There are various sources of risk of medical products that can result in injury or death, including known side effects, medication and device error, product defects, and other remaining uncertainties. These risks are influenced by many factors, including patient characteristics (such as demographics, comorbidities, concomitant medications, and health service utilization), health system factors (such as utilization practice and provider behavior), and other environmental sources. Discovery of how treatment effects vary by baseline risk is one of the important contributions of post-marketing surveillance of drugs[206]. The current measures of risk include clinical trials, spontaneous adverse event reporting systems, epidemiologic studies, and registries. Active surveillance offers the opportunity for the systematic use of observational healthcare databases, such as administrative claims and electronic health records, to improve our measures of the sources of risks. Analyses against these data can account for the measurable influents of risk to provide robust, supplemental information that can be used to both identify and evaluate potential drug safety issues. While evaluation studies have been common practice for decades, use of these data in a formal exploratory analytic framework is new and requires further research to determine its relative contribution to such a system.

When considering drug safety in a causal inference framework, one can consider Hill's considerations of 1) strength, 2) consistency, 3) specificity, 4) temporality, 5) biologic gradient, 6) plausibility, 7) coherence, 8) experimental evidence, and 9) analogy[54]. The strength of association should be considered because stronger associations may be more compelling, but weak associations do not rule out causal connections[207]. Consistency refers to the repeated observation of an association in different populations under different circumstances. Specificity relates to the number of causes that lead to a specific effect, and the number of effects produced from a given cause. Temporality refers to the necessity that the cause precedes the effect. Biologic gradient addresses the degree to which there is a dose-response relationship, where the amount of response increased with increased exposure. Plausibility reflects the scientific rationale for the existence of an association, typically in drug safety, related to the mechanism of action and the biologic pathways that lead to the effect. Coherence is the degree to which the interpretation of the association does not conflict with the current understanding of the natural history and biology of the disease. Experimental evidence for drug safety analyses typically refers to evidence that comes from human randomized clinical trials, but can also include randomized pre-clinical experiments in animal models.

An active drug safety surveillance system can apply Hill's considerations as part of its process for generating hypotheses. Specifically, analyses conducted across a network of observational databases can be used to identify potential drug safety issues based on strength, consistency, specificity, and temporality. Specifically, methods produce estimates of the strength of temporal associations between exposure and subsequent outcomes. Applying the methods to multiple sources provides an assessment of consistency, as formal tests for heterogeneity can be used to measure differences between source populations. Evaluating multiple outcomes for each drug and multiple exposures for each outcome can provide insights into the specificity of any specific drug-outcome association. However, these exploratory analysis results will not be sufficient to address issues of biologic plausibility, and the use of observational data does not meet the same standards of evidence that come from a randomized experimental design. Methods for studying dose effects requires further research, as the degree to which dose and amount of exposure can be accurately measured and used within a hypothesis generating framework is undetermined.

While hypothesis generating analyses are inherently exploratory in nature, basic principles of formal evaluation can be applied to raise the collective confidence in the reliability of the process. Research questions and statistical analyses should be specified in advance, with all methodological considerations addressed during study planning rather than after study completion. This includes decisions around definitions of exposure and outcome, inclusion/exclusion criteria imposed on the sample, and strategies for statistical adjustment[150]. Analysis processes should be fully transparent and reproducible, and should minimize subjective assessment to improve the generalizability of the approach. Many of these principles are well-defined in guidelines for conducting full evaluation studies[13, 208-210] but have not yet been adopted for exploratory analyses. With these principles in place, hypothesis generation can play an important role in an active surveillance system's contribution to causal inference of drug safety issues. These exploratory analyses can identify and prioritize areas that warrant further examination. Evaluation studies may be used to refine estimates of the strength of the association, but attention can particularly focus on biologic plausibility and coherence to put the preliminary results in proper clinical context with other evidence, including clinical trials, pre-clinical data, spontaneous adverse events, and other epidemiologic studies.

# CHAPTER THREE: METHODS

## 3.1 Overview

This study is a methodological experiment to evaluate the performance of a novel analysis technique for active drug safety surveillance. The analytical approach, called COMParator-Adjusted Safety Surveillance (COMPASS), is described (section 3.2). The evaluation of COMPASS was conducted across five observational data sources (described in section 3.3) by exploring the method's ability to identify known safety issues associated with ACE inhibitors. The experimental design, including the selection of the sample test cases of true adverse reactions and negative controls for the drug class and individual ingredients is highlighted in section 3.4. The performance measures used to assess COMPASS performance are discussed in section 3.5. The remainder of the chapter provides specific analyses conducted to support the following aims:

**Aim 1: Characterize the performance of COMPASS in identifying known safety issues associated with ACE inhibitor exposure within an administrative claims database**

**Aim 2: Evaluate consistency of COMPASS estimates across a network of disparate databases**

**Aim 3: Explore differential effects across ingredients within ACE inhibitor class**

### 3.2 COMPASS

COMParator-Adjusted Safety Surveillance (COMPASS) is a statistical algorithm that estimates adjusted rate differences and relative risks for all outcomes of interest for a given medical product through propensity score stratification across exposed and unexposed cohorts. COMPASS applies an automated heuristic for defining a comparator group based on the indication of the medical product, and provides multivariate adjustment based on key influents of risk, including person demographics, comorbidity, and health service utilization. COMPASS is not intended to be a final solution for active surveillance, but instead a first-pass screening tool to serve as a potential guide for identifying and prioritizing potential drug effects that may warrant further evaluation.

Figure 5 highlights the conceptual model that serves as the basis for COMPASS. The fundamental goal of a drug safety analysis is to assess the temporal relationship between treatment and outcome. However, in the context of an active surveillance system that leverages observational databases in a non-experimental design, specific attention is needed to minimize bias when estimating the drug-outcome association. COMPASS applies a retrospective cohort design to compare the effects of the target drug of interest to an unexposed population, defined as those exposed to an alternative treatment for the same indication. The COMPASS model focuses on minimizing bias from four primary sources: personal demographics (such as age and gender), confounding by indication, effects of comorbidity, and health serve utilization.

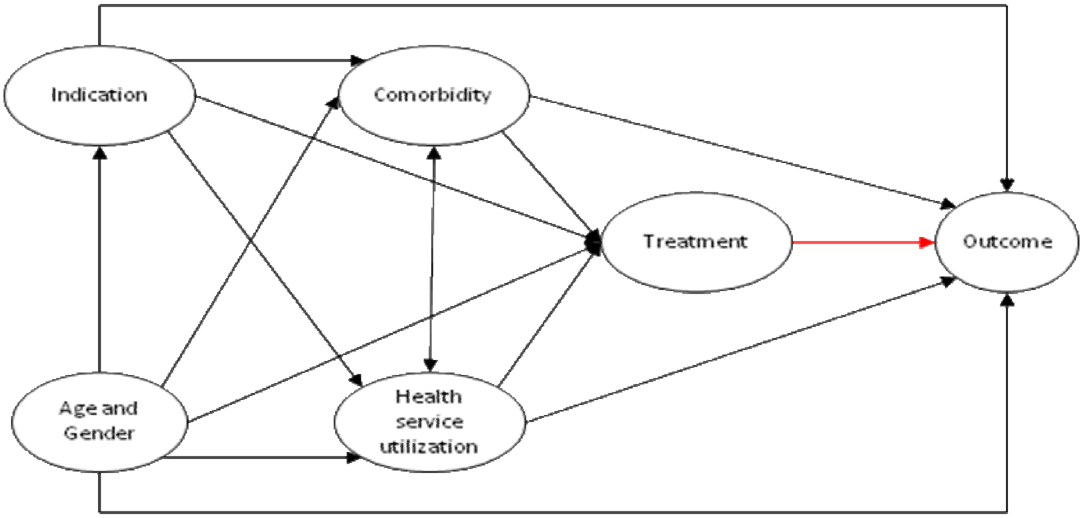


**Figure 5: COMPASS conceptual model**

The COMPASS approach incorporates several notable features into its analysis that bear particular consideration. First, it leverages large biomedical ontologies to automate comparator selection based on the indications and therapeutic classes of the target drug of interest. Second, it imposes automated study design heuristics, including cohort exclusion criteria based on contraindications and covariate selection based on FDA-approved indications and off-label uses. Third, the use of a comorbidity index and multiple measures of health service utilization as additional aggregate covariates allows for improved balancing of exposed and unexposed cohorts that are universally applicable for all outcomes while minimizing concerns of inflating bias due to unconfounded relationships with any specific outcome. Fourth, the algorithm simultaneously applies multiple risk windows to identify effects with differential time-to-event relationships, such as acute, subacute, insidious or delayed onset. Fifth, COMPASS produces a composite score based on adjusted risk differences and ratios that enables prioritization across multiple potential safety concerns based on both magnitude of effect and public health significance. Finally, in contrast to traditional pharmacoepidemiology evaluation designs, which are typically implemented to estimate the effect of one drug-condition pair, the COMPASS model is designed to be scalable to allow estimation of multiple drug-outcome pairs concurrently, and has demonstrated to be computationally feasible to screen thousands of potential adverse events within hours. This efficiency enables key principles of pharmacoepidemiology to be brought to bear during the initial exploratory phase of hypothesis generation to complement existing evaluation studies that test hypotheses once identified.

### 3.2.1 COMPASS comparator selection

COMPASS leverages the standardized vocabulary made available through the Observational Medical Outcomes Partnership (OMOP). The Standard Vocabulary contains all of the code sets, terminologies, vocabularies, nomenclatures, lexicons, thesauri, ontologies, taxonomies, classifications, abstractions, and other such data that are required for: 1) creating the transformed (i.e., standardized) data from the raw data sets; 2) searching and querying the transformed data, and browsing and navigating the hierarchies of classes and abstractions inherent in the transformed data; and 3) interpreting the meanings of the data[211].

Within OMOP, the primary use of the vocabulary has been to translate source codes into standard concepts. For example, across the OMOP data community, conditions are coded using several different coding schemes, such as ICD9, SNOMED, MedDRA, Read, and OXMIS, but the vocabulary allows all sources to be standardized into a common vocabulary (either SNOMED or MedDRA). Similarly for drugs, many source capture prescriptions using NDC, GPI, VA Product codes, or Multilex, but these codes have been mapped to RxNorm. The standard vocabulary also contains classification systems associated with its standards. For example, MedDRA provides a hierarchical structure of ‘is-a’ parent-child relationships whereby Preferred Terms (PT) are children of High Level Terms (HLT), which are children of High Level Group Terms (HLGT), which are children of System Organ Classes (SOC). The OMOP standard vocabulary offers several classifications for medical products. For example, RxNorm concepts are mapped into the National Drug File, Reference Terminology (NDF-RT), which provides classifications for mechanism of action, physiological effect, chemical structure, and indication. Notably, RxNorm is also mapped to the Anatomical Therapeutic Chemical (ATC) classification maintained by the World Health Organization (WHO) Collaborating Centre for Drug Statistics Methodology, and the National Drug Data File Plus (NDDF Plus) maintained by First DataBank. NDDF Plus provides multiple classifications for medical products, including FDA-approved indications, off-label uses, and contraindications. NDDF Plus is actively used in clinical design support tools for informing clinicians about medical information during prescription order entry. However, we are not aware of its prior use in population-level exploratory analysis of drug safety issues across observational healthcare databases.

COMPASS uses four attributes- therapeutic class, FDA-approved indications, off-label uses, and contraindications- as part of its automated heuristics, as shown in Figure 6. This graphic shows that ingredients can be mapped to each of these four attributes. It is worth highlighting that these attributes are actually mapped through RxNorm clinical drugs (which are concepts that uniquely identify product name and dose), but since active surveillance analyses are initially anticipated to be conducted at the generic ingredient-level, without immediate exploration of dose effects, the attributes have been propagated up to the ingredient level.

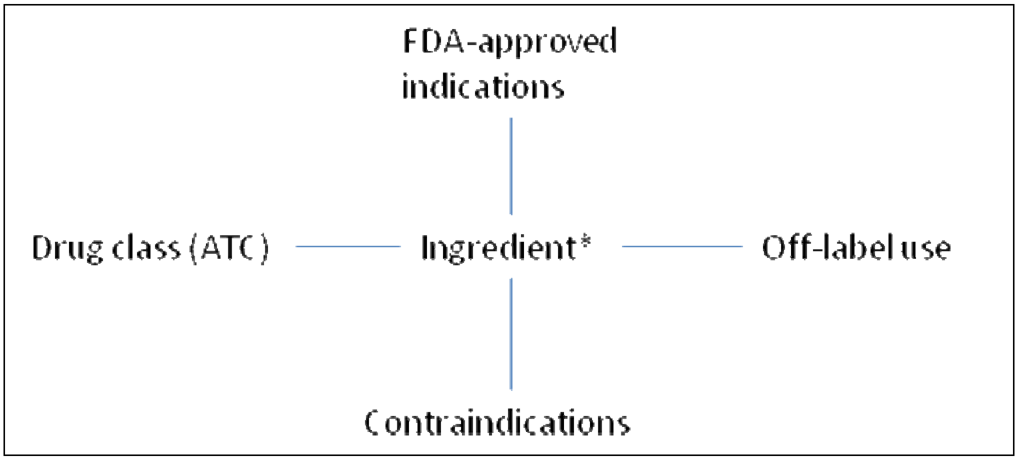


**Figure 6: Attributes of medical products used in COMPASS automated heuristics**
***Ingredient maps to these concept through RxNorm clinical drug, which contains product name and dose**

An example of how these attributes are instantiated for a given medical product, lisinopril, is shown in Figure 7. All attributes have a many-to-many relationship with ingredients, meaning that each medical product can have one or more drug classes (here, lisinopril has only one, ACE inhibitors), one or more FDA-approved indications (lisinopril has three), one or more off-label uses (lisinopril has seven in total), and one or more contraindications (lisinopril has 40).

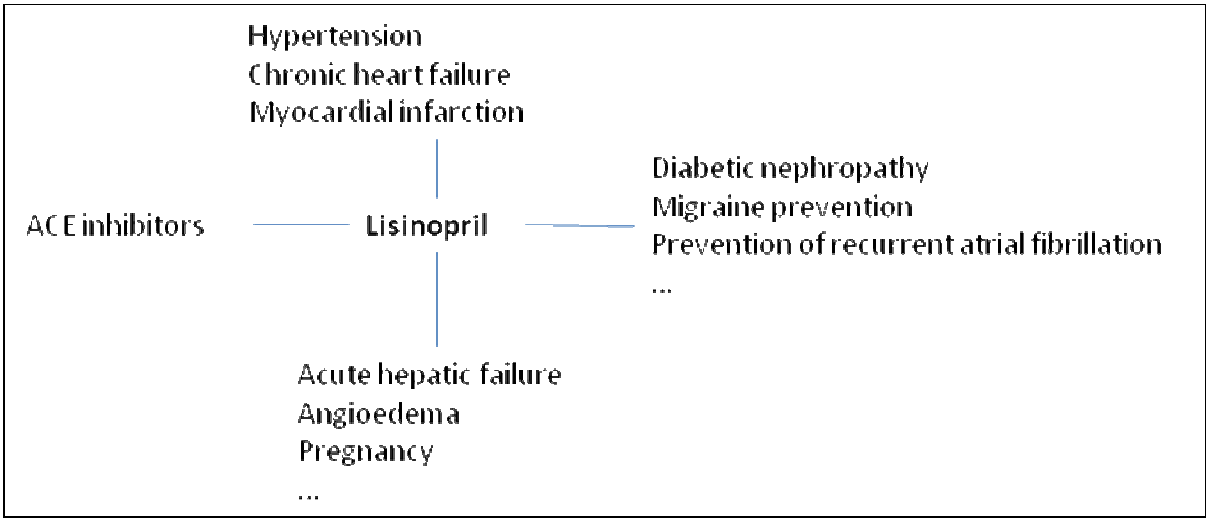


**Figure 7: Example attributes for lisinopril**

COMPASS uses these attributes to create automated heuristics for comparator selection, cohort restriction, and covariate adjustment. The logic for comparator selection is illustrated in Figure 8. The comparator group is initially defined by exposure to any medical products that have at least one of the same indications as the target drug of interest but don't share a therapeutic class. To continue with lisinopril as a working example, the algorithm identifies all drugs that have an FDA-approved indication of either 'hypertension', 'chronic heart failure' or 'myocardial infarction'. The drugs identified include ingredients from multiple drug classes, including: Angiotensin II Receptor Blockers (ARBs), such as losartan, valsartan, and candesartan; Beta Blockers, such as atenolol, metoprolol and acebutolol; Calcium Channel Blockers, such as amlodipine, nifedipine, and isradipine; diuretics, such as furosemide, amiloride, and hydrochlorthiazide; and other ACE inhibitors, such as enalapril, ramipril and captopril. The list is then restricted to those products who do not share a same therapeutic class. So, the other ACE inhibitors- enalapril, ramipril, captorpil- are removed from the indication drug list. Special consideration of combination products is taken to ensure ingredients that could be shared within the target drug are not erroneously included in the comparator drug list. As such, hydrochlorothiazide is also removed from the comparator drug list for lisinopril because the combination of the two products is marketed (brand name: Zestoretic). The final list of comparator drugs reflect a set of alternative medicines that a patient could have been prescribed by a provider for at least one of the indications that the target drug is used for. Because most observational databases do not provide explicit patient-level information that ties diagnosis to prescriptions, pharmacoepidemiology studies attempting to exploit the drug-indication relationship often do so by either assuming, inferring, or defining by explicit inclusion criteria. Moreover, pharmacoepidemiology studies commonly select a comparator drug for the unexposed cohort based on subjective assessment and clinical expertise. One reason for this approach is to minimize risk of immortal time bias that could be introduced if the unexposed population were defined by persons without any exposure (rather than an active alternative treatment). The COMPASS comparator selection heuristic provides an objective tool to construct a referent group to serve as the 'unexposed' population to compare with those exposed to the target drug of interest, and minimizes the potential bias introduced by subjective selection of only one 'similar' drug or class as an alternative treatment. The comparator selected varies by the drug under study as a proxy for 'standard of care' but does not reflect the notion of a 'no treatment' comparator group.

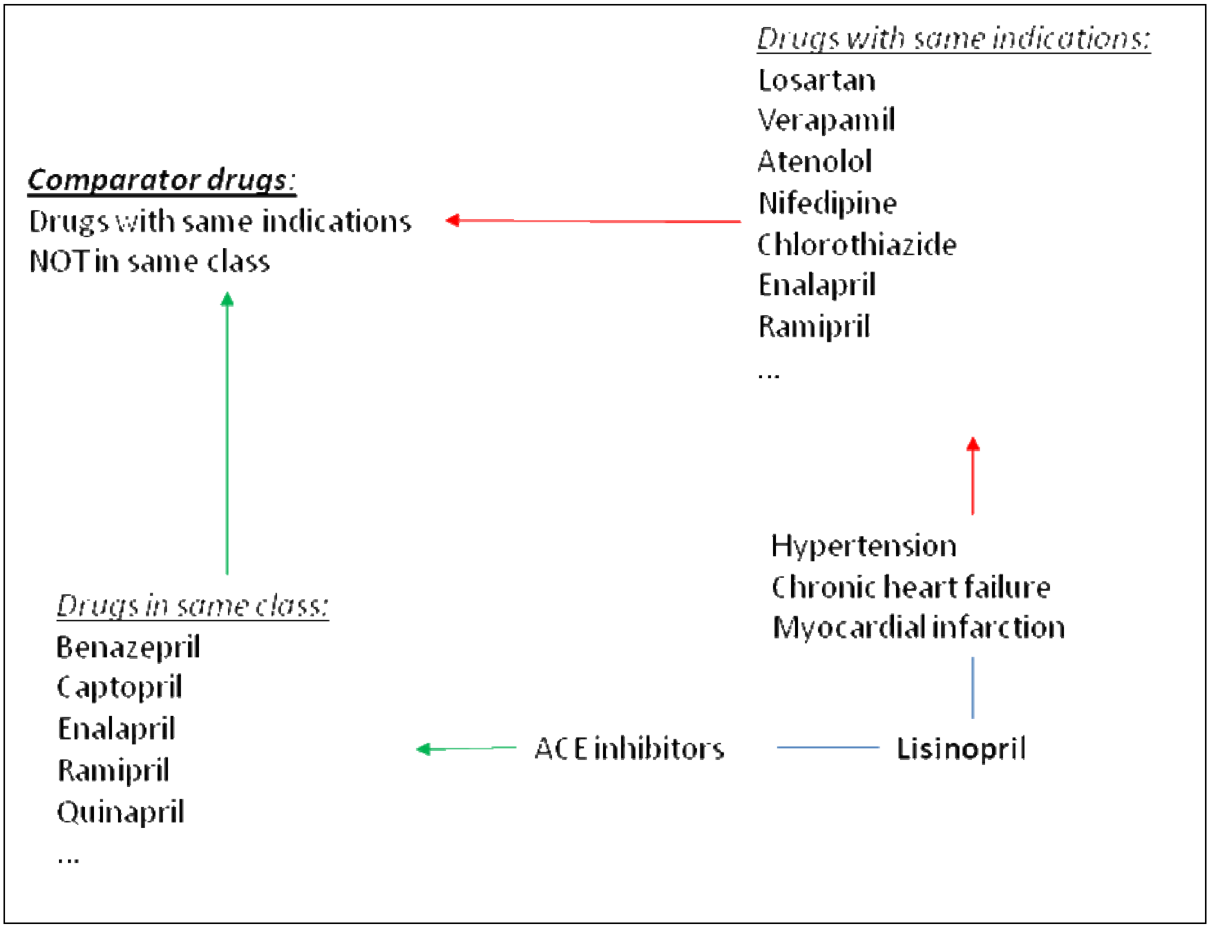
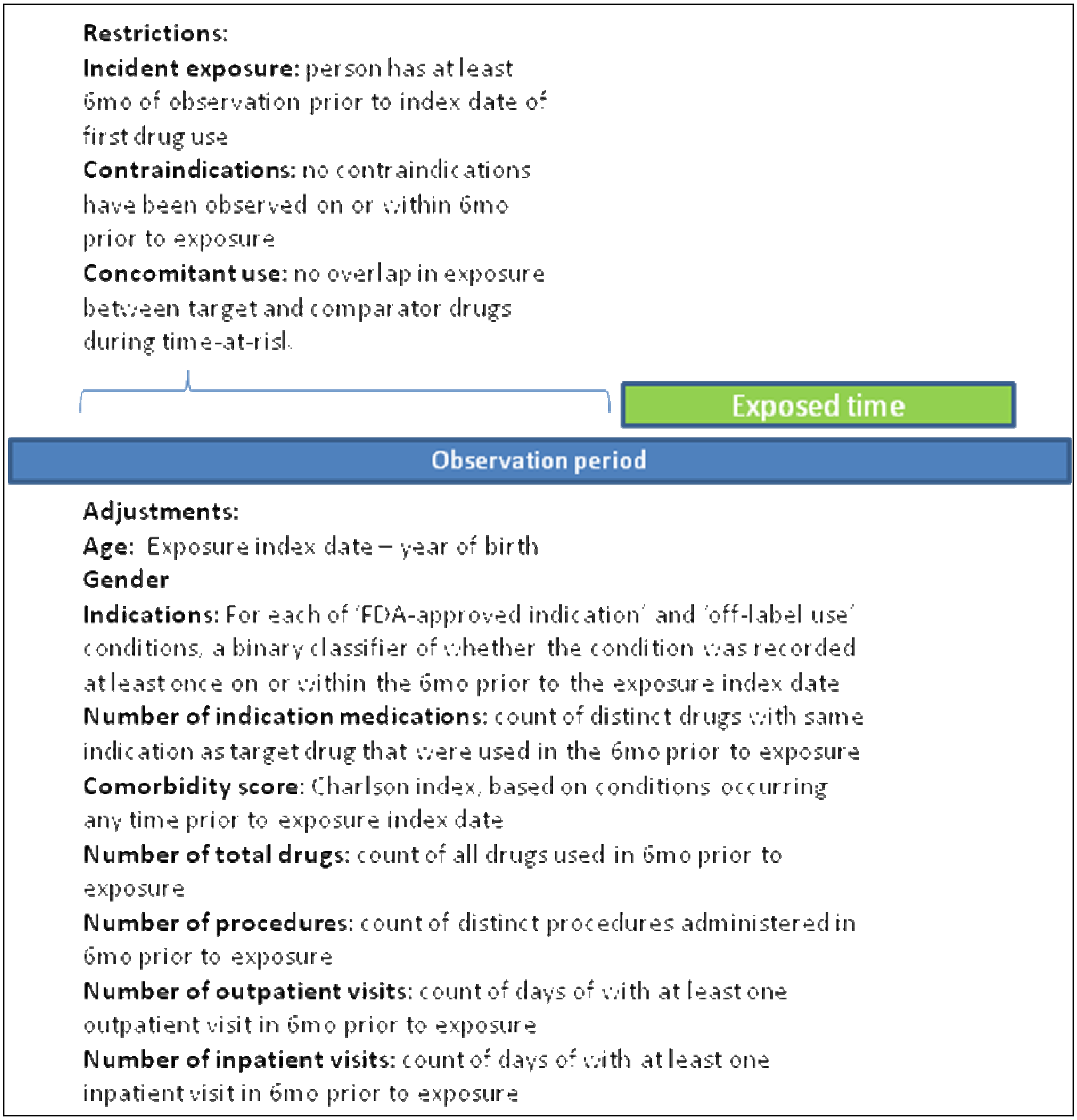


**Figure 8: COMPASS automated comparator selection heuristic**

### 3.2.2 COMPASS cohort restrictions and adjustments

While appropriate comparator selection is a critical component of the cohort design, several additional design and analysis considerations are required to improve the validity of the estimate of the drug-condition relationship. COMPASS applies a series of exclusion criteria as part of its study design and also attempts to balance the cohorts using propensity score stratification across a series of covariates. Figure 9 highlights the restrictions and adjustments imposed as part of the COMPASS analysis process.



**Restrictions:**

**Incident exposure:** person has at least 6mo of observation prior to index date of first drug use

**Contraindications:** no contraindications have been observed on or within 6mo prior to exposure

**Concomitant use:** no overlap in exposure between target and comparator drugs during time-at-risk

**Adjustments:**

**Age:** Exposure index date – year of birth

**Gender**

**Indications:** For each of 'FDA-approved indication' and 'off-label use' conditions, a binary classifier of whether the condition was recorded at least once on or within the 6mo prior to the exposure index date

**Number of indication medications:** count of distinct drugs with same indication as target drug that were used in the 6mo prior to exposure

**Comorbidity score:** Charlson index, based on conditions occurring any time prior to exposure index date

**Number of total drugs:** count of all drugs used in 6mo prior to exposure

**Number of procedures:** count of distinct procedures administered in 6mo prior to exposure

**Number of outpatient visits:** count of days of with at least one outpatient visit in 6mo prior to exposure

**Number of inpatient visits:** count of days of with at least one inpatient visit in 6mo prior to exposure

**Figure 9: COMPASS pre-exposure design considerations**

COMPASS applies an incident user design to compare new users of alternative treatments. Incident use is inferred by requiring that persons have at least 6 months of observation prior to the index date of the first drug use of either the target drug or comparator. It is possible that patients could have been exposed previously, but that exposure was not observed due to the period of data capture contained within the observational database. It is assumed that any potential ‘prevalent use’ that occurs due to lack of data coverage represents a small but non-differential bias, and sensitivity analyses can be conducted to evaluate the robustness of any findings by varying the length of the washout period to be more or less than 6 months. Because of the incident use providing a comparable initiation of treatment between cohorts, it has been argued that the populations are more likely to be similar in characteristics that might not be observable in the database[90]. Restricting prevalent use allows for a clear temporal sequence for confounder adjustment while minimizing concern of adjusting for intermediate consequence of treatment rather than just treatment predictors[90]. Definition of treatment initiation also allows for a more precise measure of time-at-risk that can be used to assess adverse events with different time-to-event relationships, such as acute and delayed onset.

An additional restriction imposed in COMPASS is that all persons with concomitant use of drugs in the exposed and unexposed lists during the time-at-risk window are excluded. This restriction ensures that events attributable to the target drug are not erroneously classified for the unexposed cohort, or vice versa. The risk window defined will influence the degree to which the concomitant use will restrict the overall sample; estimating potential acute onset events, where only the first 30 days following exposure start are of interest, will be less restrictive than exploration of insidious events, where all time exposed needs to be non-overlapping.

Another potential source of bias introduced by the automated comparator selection heuristic is the potential for factors that influence treatment avoidance. A classic example of channeling bias is studying gastrointestinal effects among users of Cox-2 inhibitors and other non-steroidal anti-inflammatory drugs (NSAIDs). One of the primary benefits of Cox-2 inhibitors relative to NSAIDs was greater GI tolerability; as a result, prescribers tended to avoid the use of traditional NSAIDs to those patients with peptic ulcers and other gastrointestinal hemorrhaging and would channel those patients to use Cox-2 inhibitors. Without adequate restriction or adjustment for the channeling effect, analyses can produce biased estimates that indicate Cox-2 inhibitor use has an increased GI risk, rather than a preventative effect. COMPASS leverages the contraindication information available in the OMOP standard vocabulary to impose automated exclusion criteria on the cohorts to minimize this potential source of bias. As shown in Figure 10, all contraindications are mapped through ICD9 codes to SNOMED clinical findings. Patients are removed from the cohort if there are one or more contraindication condition era records that start in the 6 months prior to the exposure index date. For the lisinopril example, patients with 'acute hepatic failure', 'angioedema', 'pregnancy', and any other listed contraindication observed in their record are excluded from both the exposed and unexposed cohorts. This restriction eliminates the subpopulation that may be more predisposed to known risks.

After all restriction criteria have been applied to the cohorts, COMPASS creates a series of covariates to use in balancing baseline characteristics to further refine the effect estimates. Balancing is achieved through propensity score stratification [45, 106, 107, 212, 213], whereby the covariates are used in a multivariate logistic regression model to estimate the probability of the person being exposed to the target drug vs. the comparator drugs, and persons in both cohorts are stratified into quantiles based on this probability. Effects are measured within these propensity score strata, and composite summaries are constructed using Mantel-Haenzsel estimator. Strata that contain only exposed or unexposed patients are excluded from analysis, as a means to ensure overlap between comparator populations.

Stratification is chosen over matching because it is computationally simpler to construct, and preserves sample without oversampling or imbalanced weighting for outlier patients[214]. In prior applications of propensity score balancing, covariates are selected either through subjective assessment and clinical expertise or through heuristics that measure the potential degree of confounding based on a variable's relationship to both treatment and outcome[108, 116]. Past studies have demonstrated that inappropriate use of non-confounded covariates that are related to treatment but not outcome can inflate variance estimates around the treatment-outcome effect[84, 108, 215]. A challenge in active surveillance, where hundreds of thousands of drug-condition pairs may warrant investigation and may need to be explored rapidly on a regular basis, is that clinical expert review is likely infeasible and computations requiring pairwise comparisons may not be scalable for use in the initial exploratory stages. As such, COMPASS creates a restricted set of covariates, based on personal demographics, treatment indication, comorbidity, and health service utilization, which are expected to address the primary sources of bias while avoiding unconfounded relationships, to provide cohort balancing that is universally sufficient to facilitate simultaneous estimates of all outcomes.

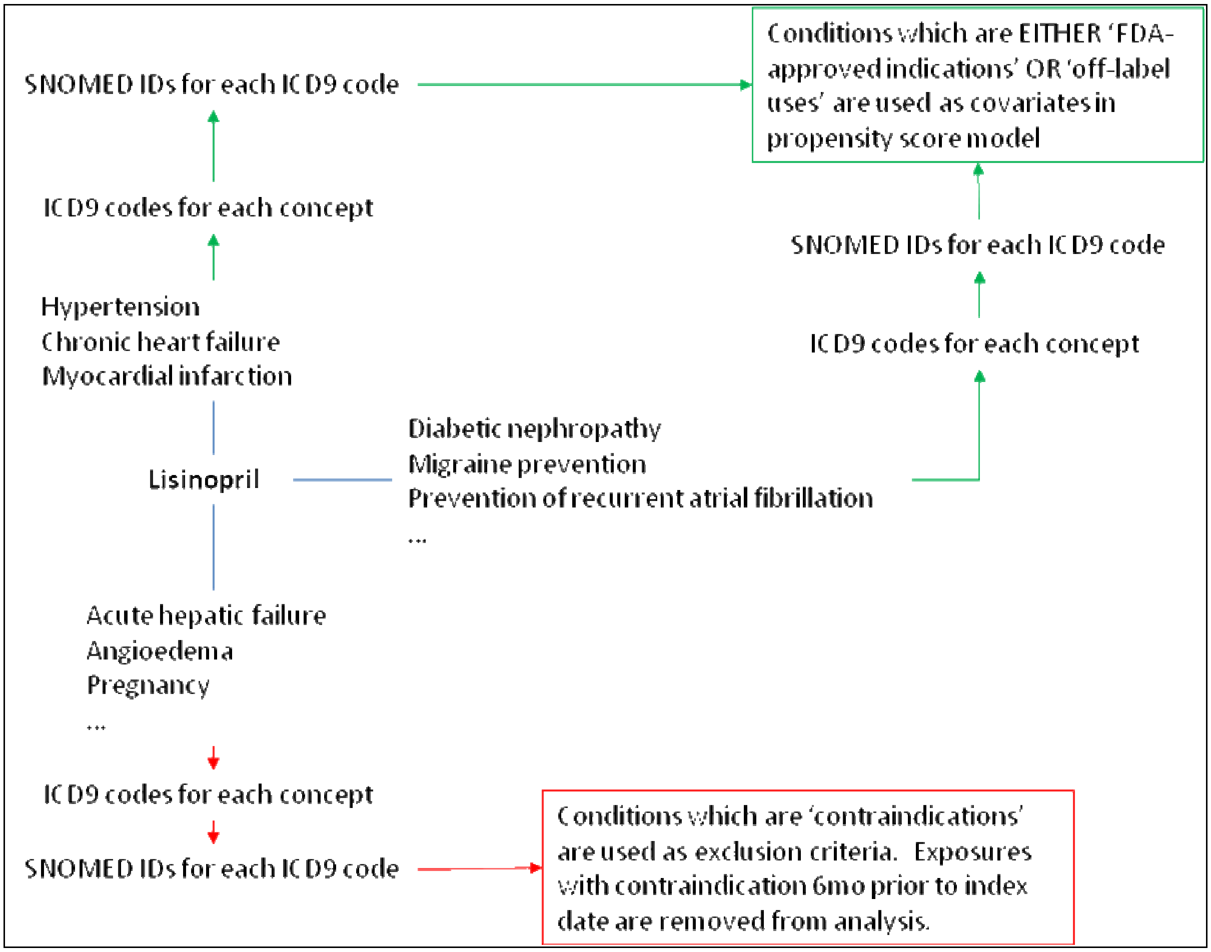

SNOMED IDs for each ICD9 code
Conditions which are EITHER 'FDA-approved indications' OR 'off-label uses' are used as covariates in propensity score model
ICD9 codes for each concept
SNOMED IDs for each ICD9 code
Hypertension
Chronic heart failure
Myocardial infarction
ICD9 codes for each concept
Lisinopril
Diabetic nephropathy
Migraine prevention
Prevention of recurrent atrial fibrillation
...
Acute hepatic failure
Angioedema
Pregnancy
...
ICD9 codes for each concept
SNOMED IDs for each ICD9 code
Conditions which are 'contraindications' are used as exclusion criteria. Exposures with contraindication 6mo prior to index date are removed from analysis.

**Figure 10: COMPASS automated design refinement process**

Covariates associated with indication are a primary consideration within COMPASS. The comparator drugs are selected based on potential for having a similar indication as the target drug. However, the observed prevalence of the indications prior to exposure is not accounted for. As such, there could be potential for imbalance between the exposed and unexposed populations, resulting in confounding by indication. For example, using the lisinopril example, if the majority of patients prescribed lisinopril use the medication for their hypertension, but the majority of patients in the unexposed population are being treated for myocardial infarction, there could be cohort differences in the cardiovascular profile of the patients that could bias comparisons in measured post-exposure effects. COMPASS attempts to address this potential concern by using these indications as covariates to be balanced through propensity score stratification prior to analysis. Specifically, COMPASS constructs binary classifiers for each medical concept identified as either an FDA-approved indication or an off-label use. The concepts are constructed through the OMOP standard vocabulary by mapping the NDDF concepts to one or more ICD9 codes, which are then mapped to one or more SNOMED clinical findings. For each indication concept, persons are classified as 1 if at least one of the SNOMED codes comprising the indication is recorded in a condition era start within the 6 months prior the exposure index date, and 0 otherwise. Figure 10 highlights the heuristic for the lisinopril example; concepts are constructed for all FDA-approved indications ('hypertension', 'chronic heart failure', and 'myocardial infarction') and all off-label uses (including 'diabetic nephropathy', 'migraine prevention', and 'prevention of recurrent atrial fibrillation') in the FirstDataBank vocabulary through the mapping via ICD9 and SNOMED.

A related effect is the number of drugs previously used for the indications. While all patients are incident users to the drug of interest, the cohort definition does not guarantee that those patients hadn't attempted other alternative treatments for their underlying disease prior to initiating treatment to the target or comparator drug. A patient receiving first-line treatment for a disease may have different characteristics than someone who has switched due to prior treatment failures. The number of prior drugs used for the indications serves as a proxy for the number of treatment switches and can potentially inform the level of underlying disease severity insofar as multiple treatments are attempted due to the inherent complexity of the disease or lack of response to initial treatments by the patient. The covariate, 'number of indication medications', is measured as the count of distinct ingredients used within the 6 months prior to the index date that share at least one indication as the target drug. In the lisinopril example, this could include the number of beta blockers, diuretics, ARBs, or other ACE inhibitors attempted in the 6 mo before lisinopril initiation. A count of 0 would be potentially indicative of a patient who is using the target drug as first-line treatment for one of the indications, while larger counts may increase the likelihood that the patient is switching to the target treatment after prior treatment attempts.

Beyond the variable set of covariates defined by the target drug attributes, COMPASS also applies a defined set of covariates that are independent of the target drug but are thought to be important in any drug safety analysis. These include: age, as measured in years by the difference in the index year from the patient's year of birth; gender, as a binary classifier indicating male or female status; the Charlson comorbidity index, as a score reflecting overall disease status, based on conditions observed prior to exposure index date[94]; and four methods of health service utilization. 'Number of drugs' is measured by the count of distinct ingredients used within the 6 months prior to the index date. 'Number of procedures' is measured as count of the distinct procedures administered within the 6 months prior to the index date. 'Number of outpatient visits' and 'number of inpatient visits' reflect the number of distinct days for which services were initiated in outpatient and inpatient centers, respectively. The 'inpatient' measure included both hospital stays and emergency room visits not requiring hospitalization.

The exposed and unexposed cohorts are stratified by propensity score $P_i$ , estimated by the following logistic regression:

$$\ln(\frac{P_i}{1-P_i}) = \alpha + \beta_{age} * age_i + \beta_{gender} * gender_i + \beta_{indication} * indication_i + \beta_{indicationdrugs} * indicationdrugs_i + \\ \beta_{comorbidity} * comorbidity_i + \beta_{inpatient} * inpatent_i + \beta_{outpatient} * outpatent_i + \beta_{drug} * drugs_i$$

The computational efficiency that makes COMPASS viable as an initial hypothesis-generating tool also comes at the sacrifice of precision of the association estimates. Specifically, global covariates (such as the comorbidity index and aggregate health service utilization measures) are used in lieu of drug- or disease-specific covariates because individual covariates could have unobserved confounding, but the confounding effects would vary by outcome. Preliminary studies using the Charlson comorbidity index in the propensity score model found improved balance not only of the index, but also reduced differences in most of the constituent comorbidities that comprise the index as well.

### 3.2.3 COMPASS risk windows

The time-to-event relationship between an exposure and an adverse event can vary based on the pharmacologic effect of the medicine and the disease progression of the event[216]. Some events, such as anaphylaxis reactions, commonly have an acute onset, and are generally observed shortly after initial exposure to the medication or never at all. Other events may have a delayed onset, such as cancer, which may result from long-term exposure. Still other events may have different time-to-event relationships based on the effect with the drug; studies exploring drug-related relationships with hip fracture have shown insidious onset with benzodiazepines due to risk of dizziness and falls, while a delayed effect observed with proton pump inhibitors hypothesized to be due to long-term calcium malabsorption.

Traditional pharmacoepidemiologic evaluation studies that focus on one drug-outcome pair typically have a hypothesis of the type of effect under study that can be used to define the study risk window. In the active surveillance paradigm, where we are exploring multiple outcomes to identify potential effects that warrant further review, we do not have the luxury of knowing the time-to-event relationship. Alternative active surveillance approaches may pre-specify a risk window of interest; for example, the use of self controlled case series would traditionally assume insidious onset since the time exposed is used to define the time-at-risk. In COMPASS, we seek an alternative to pre-specifying the time-to-event relationship by instead concurrently testing four clinical scenarios (see Figure 11). Time-at-risk windows are defined as either acute, subacute, insidious, or delayed onset. The acute onset window captures all events that occur within the 30 days following exposure start. Subacute onset includes all events within 60 days of treatment initiation, subsuming the acute risk window. Insidious onset is defined as any time during exposure (from exposure start to exposure end) or within 30 days following exposure end. The additional 30 day surveillance window is to accommodate misclassification in exposure end date estimation, and to capture events that may proceed a patient ceasing treatment and seeking an alternative therapy (potentially due to lack of effectiveness or tolerability due to side effects). Delayed onset is defined by the period from 180 days following exposure start until the end of the observation period. The delayed window may or may not include period of exposure. All risk windows serve as intent-to-treat analyses and are right-censored by the observation period end date. It is important to highlight that, as with any intent-to-treat analysis, this approach may be susceptible to selection bias due to treatment stopping, switching, augmentation, and non-adherence.

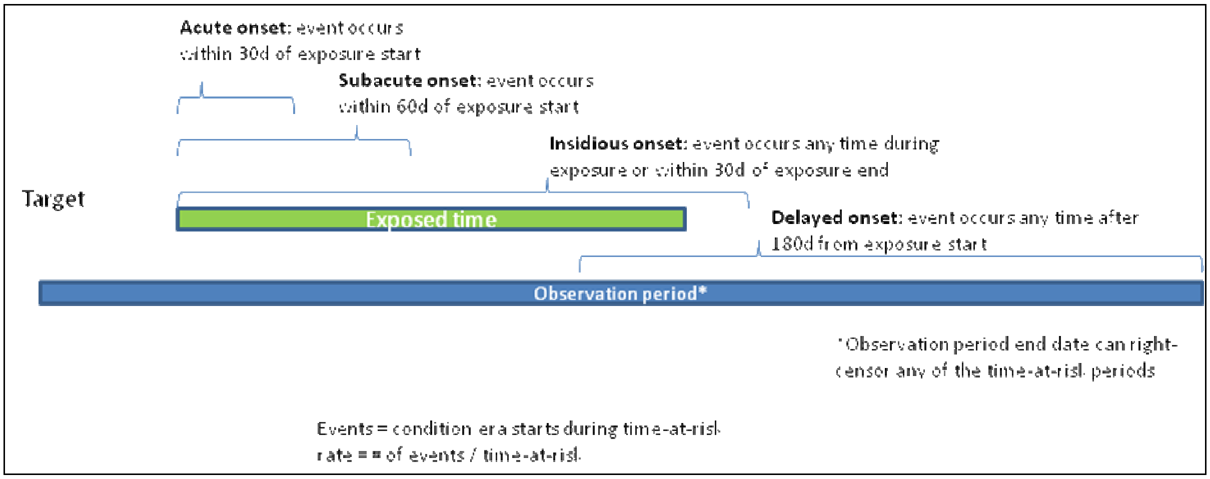

Acute onset: event occurs within 30d of exposure start
Subacute onset: event occurs within 60d of exposure start
Insidious onset: event occurs any time during exposure or within 30d of exposure end
Target
Exposed time
Delayed onset: event occurs any time after 180d from exposure start
Observation period*
*Observation period end date can right-censor any of the time-at-risk periods
Events = condition era starts during time-at-risk
rate = # of events / time-at-risk

**Figure 11: COMPASS alternative risk windows**

Within each risk window, events are identified by condition era start dates that fall within the time-at-risk. When evaluating prevalent events, each person can contribute one or more events during exposure. Across the cohort population, the event rate is estimated as the number of events / total time-at-risk. For each risk window, adjusted rate ratios (ARR) and adjusted rate differences (ARD) between cohorts can be estimated for each outcome as:

$$ARR_o = \frac{\sum_s w_s \sum_i (outcome_{i,o,s,\exp osed} / timeatrisk_{i,o.s,\exp osed})}{\sum_s w_s} / \frac{\sum_s w_s \sum_i (outcome_{i,o,s,un\exp osed} / timeatrisk_{i,o.s,un\exp osed})}{\sum_s w_s}$$

$$ARD_o = \frac{\sum_s w_s \sum_i (outcome_{i,o,s,exposed} / timeatrisk_{i,o.s,exposed})}{\sum_s w_s} - \frac{\sum_s w_s \sum_i (outcome_{i,o,s,unexposed} / timeatrisk_{i,o.s,unexposed})}{\sum_s w_s}$$

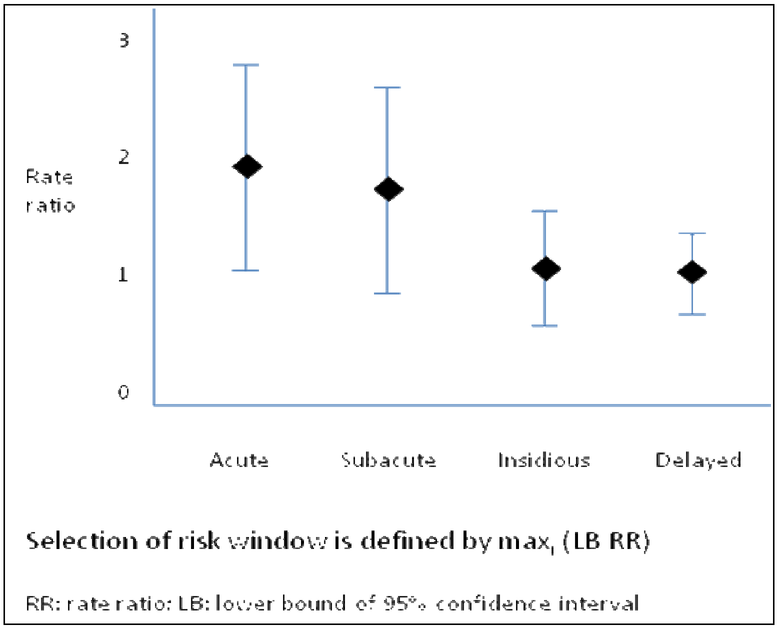
$$\text{var}(ARD_o) = \text{var}(AR_{o,\exp osed}) + \text{var}(AR_{o,un\exp osed})$$

$$\text{var}(AR_{o,t}) = \frac{\sum w_s^{2} (outcome_{o,s,t} / timeatrisk_{o.s,t}^{2})}{(\sum_s w_s)^2}$$

$$ARDLB95_o = ARD_o - 1.96 * \sqrt{\text{var}(ARD_o)}$$

where o is the outcome, s in S is the propensity score strata derived from $P_i$, w are the weights for each strata based on the inverse of the variance, and i is the index for each person.

Within an outcome, the selection of the risk window is made by identifying the risk window that yields the maximum ARRLB95 (see Figure 12). This selection criteria aims to prioritize the risk window that has the largest relative effect. The lower bound is used in lieu of the point estimate to filter out unstable estimates generated by small outcome counts.



Selection of risk window is defined by $\max_i$ (LB RR)

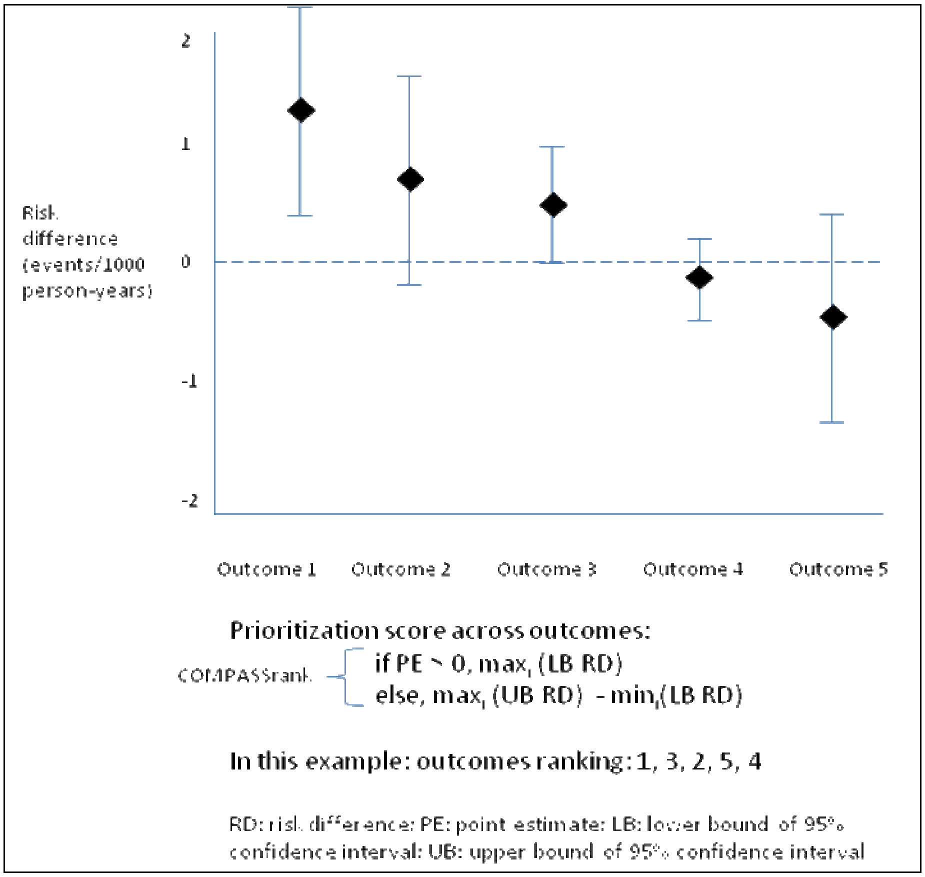
RR: rate ratio; LB: lower bound of 95% confidence interval

**Figure 12: COMPASS prioritization across risk windows**

### 3.2.4 COMPASS prioritization score

Both the adjusted rate ratio and rate difference measures provide useful information for assessing the significance and potential public health impact of a particular drug-outcome pair. However, in the context of active surveillance, where a potential aim is to explore multiple drugs and conditions over time, there is a further need to prioritize the observed effects so that the limited resources available can focus on those drug-outcome pairs that are more likely to be true causal relationships that warrant some sort of intervention. One conventional approach to prioritization is to construct a dichotomous threshold, whereby pairs with a score that exceeds the threshold are considered 'signals' and those not meeting the threshold are not evaluated further. In the context of the outputs available from COMPASS, one could derive a signal threshold based on the rate ratio, the rate difference, the confidence intervals around those point estimates, or some combination therein. For example, drug-outcome pairs with ARRLB95 > 1 reflect pairs that have statistically significant rate ratios that indicate some increase in risk between the target and comparator cohorts. Alternatively a threshold of ARR > 2 reflects the estimated effect of the target drug is more than double that of the unexposed population; note, there is no inherent measure of uncertainty embedded in the decision making. Thresholds based on rate ratios emphasize magnitude of effect size, but do not characterize proportion of population effects. In contrast, thresholds based on adjusted rate differences, such as ARD > 1/1000, can provide a designation of a level of potential public health significance that is required to warrant investigation. Composite thresholds can impose further restrictions, such as ARR > 2 and ARD > 1/1000, which would only be satisfied for those drug-outcome pairs with a large effect size and a large potential public health impact. When applying thresholds, performance can be clearly classified by the degree to which the threshold generates true positives, false positives, true negatives, and false negatives. However, method performance becomes a function of the threshold, the dichotomization makes prioritization amongst the ‘signals’ and ‘non signals’ difficult, and it is not clear what the appropriate threshold to set for any given drug or outcome.

As an alternative approach, COMPASS constructs a prioritization score that allows for rank-ordering drug-outcome pairs based on a single scalar value. The score is based on the confidence interval of the risk difference; if the point estimate of the ARD > 0, then the score is the lower bound of the 95% confidence interval, otherwise when the ARD < 0, the score is the upper bound of the 95% confidence interval – the minimum ARDLB across all outcomes. Essentially, all outcomes with positive point estimates are prioritized above outcomes with negative point estimates, but amongst positive effects, the largest lower bound is prioritized to reflect the highest confidence in a true association, and amongst the negative effects, the largest upper bound is prioritized to reflect the lowest confidence in a lack of association. Figure 13 provides an illustration of how the prioritization score would rank 5 outcomes. Outcome 1 has the largest overall ARD and ARDLB so is ranked first, but outcome 2 is prioritized over outcome 3 despite having a lower ARD because of its higher ARDLB. Outcomes 4 and 5 are deprioritized from the first three because ARD < 0, but outcome 5 is ranked higher than 4 because the upper bound reflects greater uncertainty that there may still be a positive effect.



Prioritization score across outcomes:

COMPASSrank = { if PE > 0, $\max_i$ (LB RD) ; else, $\max_i$ (UB RD) - $\min_i$(LB RD) }

In this example: outcomes ranking: 1, 3, 2, 5, 4

RD: risk difference; PE: point estimate; LB: lower bound of 95% confidence interval; UB: upper bound of 95% confidence interval

**Figure 13: COMPASS prioritization across outcomes**

### 3.2.5 COMPASS summary

COMPASS is intended to be a fully-automated exploratory analysis method that allows safety scientists to specify a drug of interest and generate hypotheses of potential drug-related adverse associations that may warrant further evaluation. COMPASS was developed using SQL and SAS 9.1 against a Oracle 11g database running in the OMOP Research Lab on a SUN M5000 server with on a Windows 2003 server with 16 (8x2) 2.14GHz CPU, 64GB RAM, and 50TB of storage. The implementation was customized for use with the OMOP common data model, but the algorithm could be generalized to the format of any observational healthcare database that would allow for the exploration of temporal relationships between drug exposure and condition occurrence.

The COMPASS algorithm is executed with the following steps:

1. INPUT: Define the target drug of interest
2. Identify the FDA-approved indications of the target drug
3. Identify the ATC drug class of the target drug
4. Define the comparator drugs as drugs with the at least one of the same indications as the target drug and not in the same ATC drug class (see Figure 8)
5. Select all persons with at least one exposure to the target drug (target cohort)
6. Select all persons with at least one exposure to one of the comparator drugs (comparator cohort)
7. Identify the contraindications of the target drug
8. Restrict cohorts to exposures with at least 6 months of observation prior to first exposure index date
9. Restrict cohorts to exposures without any contraindication conditions within 6 months prior to first exposure index date (see Figure 10)
10. Exclude persons who have overlapping exposure to target and comparator drugs during the time-at-risk
11. Identify the off-label uses of the target drug
12. Create covariates (see Figure 9)
13. Calculate propensity score (probability of being in target cohort) from all covariates

14. Stratify population into 20 quantiles based on rank-ordered propensity scores. No persons from the target or comparator cohorts are dropped from analysis, but balancing statistics are produced to allow scientists to review if there is sufficient sample in each quantile for each cohort.
15. For all outcomes, calculate the strata-specific incidence rate for each of four risk windows, based on acute, subacute, insidious and delayed onset (see Figure 11). Both cohorts are right-censored at time of observation period end.
16. Calculate adjusted rates, adjusted rate differences and adjusted rate ratios with associated confidence intervals for all outcomes and all risk windows.
17. For all outcomes, select the risk window with the maximum rate ratio (see Figure 12).
18. Calculate COMPASS prioritization score based on rate difference across outcomes (see Figure 12Figure 13).
19. OUTPUT: COMPASS prioritization scores for all outcomes, rate ratio and rate difference estimates for all risk windows, cohort propensity score balance statistics

### 3.3 Data Sources

The primary data source under study is the Thomson Reuters MarketScan Commercial Claims and Encounters (CCAE), a large administrative claims database containing 59 million privately insured lives. CCAE provides patient-level de-identified data from inpatient and outpatient visits and pharmacy claims of multiple insurance plans. In addition to CCAE, the performance of COMPASS was evaluated against the MarketScan Lab Database (MSLR), MarketScan Medicaid Multi-State Database (MDCD), MarketScan Medicare Supplemental and Coordination of Benefits Database (MDCR), and the GE

Centricity electronic health record (GE). MSLR contains 1.5 million persons representing a largely privately-insured population, with administrative claims from inpatient, outpatient, and pharmacy services supplemented by laboratory results. MDCD provides administrative claims data for 11 million Medicaid enrollees from multiple states. MDCR captures administrative claims for 5 million retirees with Medicare supplemental insurance paid for by employers, including services provided under the Medicare-covered payment, employer-paid portion, and any out-of-pocket expenses. GE contains patient-level data for 11 million persons captured at the point of care from a consortium of providers using the GE Centricity electronic health record system in their outpatient and specialty practices. Table 1 provides a comparison of the source populations and data availability.

The five sources reflect the broad diversity of data available and under consideration for a national active surveillance system. They include various populations of interest with different demographics and health behaviors (privately insured, Medicaid young, Medicare elderly) as well as both primary data capture processes (administrative claims and electronic health records). The diversity in the source populations is likely to significantly influence active surveillance methods performance, though the potential effect has not been previously empirically measured. In particular, the Medicare database reflects an older population with higher drug use, more comorbidities, and greater health service utilization than any other database, so can be expected to potentially reflect a higher-risk population that is also more predisposed to confounded relationship between exposure and outcome. The first and third aim focused on CCAE since it is the largest database and is most representative of the general population. Aim 2 applied COMPASS to all sources to study how the underlying data can influence method performance in identifying drug safety issues.

**Table 1: Source population characteristics**

| | | CCAE | MSLR | MDCD | MDCR | GE |
|---|---|---|---|---|---|---|
| Population (N) | | N=59,836,290 | N=1,466,617 | N=11,188,360 | N=4,655,736 | N=11,216,208 |
| | Years of coverage | 2003-2008 | 2003-2008 | 2002-2007 | 2003-2008 | 2000-2008 |
| Gender | | | | | | |
| | Male: N (%) | 29,173,105 (48.75) | 515,174 (35.13) | 4,665,014 (41.70) | 2,071,968 (44.50) | 4,751,444 (42.36) |
| | Female: N (%) | 30,663,185 (51.25) | 951,443 (64.87) | 6,523,346 (58.30) | 2,583,768 (55.50) | 6,460,828 (57.60) |
| Age (yrs) | | | | | | |
| | Mean (SD) | 32.4 (18.1) | 39.1 (17.5) | 23.4 (22.7) | 74.5 (8.0) | 40.6 (22.0) |
| Observation period length (mo) | | | | | | |
| | Mean (SD) | 21.2 (18.6) | 18.7 (11.1) | 14.2 (13.8) | 31.9 (22.9) | 24.0 (31.3) |
| Number of drug exposure records per person | | | | | | |
| | Median (25-75 %tile) | 9 (3-28) | 14 (5-35) | 14 (5-38) | 60 (20-134) | 8 (3-22) |
| Number of condition occurrence records per person | | | | | | |
| | Median (25-75 %tile) | 15 (5-39) | 27 (12-56) | 24 (9-63) | 57 (20-129) | 5 (2-10) |
| Number of procedure occurrence records per person | | | | | | |
| | Median (25-75 %tile) | 20 (7-52) | 39 (19-77) | 31 (12-70) | 72 (26-154) | 10 (3-24) |

CCAE: Thomson MarketScan Commercial Claims and Encounters; MSLR: Thomson MarketScan Lab; MDCD: MarketScan Medicaid Multi-State Database; MDCR: MarketScan Medicare Supplemental and Coordination of Benefits Database; GE: GE Centricity electronic health record; SD: standard deviation

All data sources have been transformed into the OMOP common data model[211, 217]. The common data model is a single data schema that can be applied to disparate data types to enable consistent and systematic application of analysis methods to produce comparable results across sources. The OMOP common data model leverages standardized terminologies to transform sources that use different coding schemes for drugs and conditions into a common vocabulary. The common data model also imposed consistent transformation rules for key data elements, such as logic for inferring drug exposure length. The model was designed to accommodate and distinguish between data elements from disparate sources, such as recording drug exposure by delineating between prescription dispensings captured by pharmacy claims, procedural administrations entered on medical claims, and prescriptions written and medication history lists recorded in electronic health records system.

Conceptually, the common data model core module has eight entities, shown in Figure 14. These are:

1. Person, which includes attributes such as gender and year of birth

2. Observation Period (the time at which health care information may be available)

3. Drug Exposure (i.e., the association between Person and Drug for a specific time period)

4. Health Outcome of Interest, which may be based on a combination of:

5. The Person's medical Condition(s)

6. One or more Clinical Observations about the Person (e.g., laboratory test results)

7. One or more Medical Procedures that the Person required

8. One or more Visits for health care services for the Person

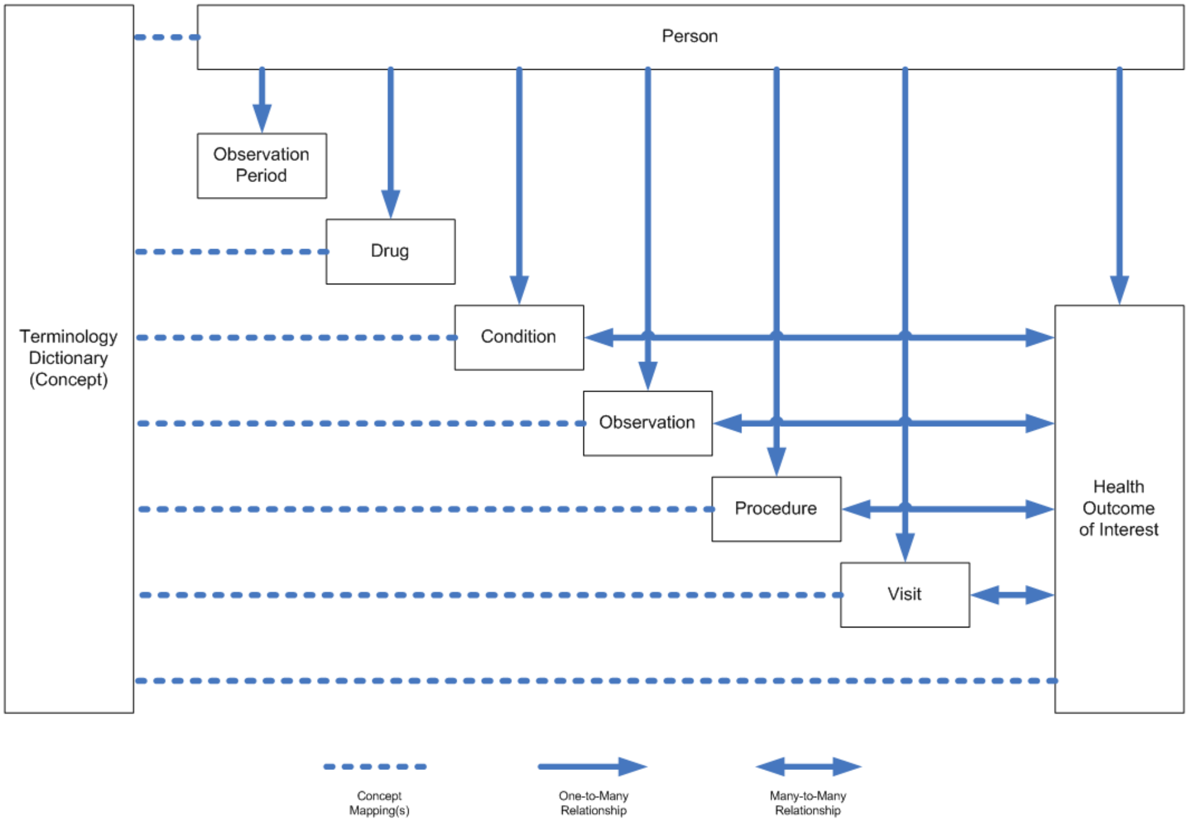


**Figure 14: OMOP Common Data Model conceptual schema**

For each source, the corresponding person-level data elements from the raw databases are transformed into each of the common data model entities. The analysis common data model is constrained to only include data elements during periods of time where a person is potentially eligible to have both exposure and outcome recorded. In the context of administrative claims, this restriction corresponds to requiring eligibility where patients have both pharmacy and medical benefit coverage. In clinical systems, these eligibility periods can be defined by the first and last observation recorded. In both cases, patients must contribute at least one valid observation period, and all data elements that fall within those periods of time are recorded in the corresponding tables. Data that falls outside a valid observation period is excluded from analysis.

The databases code drug utilization using several source terminologies, including National Drug Code (NDC) and Generic Product Identifier (GPI), as well as procedural administrations coded in Current Procedural Terminology, 4th edition (CPT-4), Healthcare Common Procedure Coding System (HCPCS), and ICD9 surgical procedure codes. These source codes have been mapped into RxNorm as the standard terminology, which provides a common classification of clinical drugs and ingredients.

As part of its design, the common data model contains a DRUG_EXPOSURE table, which stores all verbatim records from the source database that could be potentially used to infer drug exposure. Most source databases provide an identifier for the medical product used and an exposure start date, which requires inferring exposure period length based on other available records. For example, this table may contain prescription dispensings (with information such as quantity and days supply), or prescriptions written with quantity of medicine (with information such as number of refills), or medication history listings (which may provide a drug stop date).

Because source databases may vary significantly in the available fields that could be used to infer exposure, a supplemental data table, DRUG_ERA, was created. The DRUG_ERA table is intended to have one common structure for maintaining periods of persistent exposure. DRUG_ERA is a derived table, based on the DRUG_EXPOSURE table that pre-processes the data to make it more analysis-friendly and minimize the computational burden. The intent behind developing this framework is to establish one systematic, transparent process for building DRUG_ERAs that can be consistently applied across all drugs in a database, and across multiple databases.

Drug era construction is a person-level data transformation that serves two purposes: 1) rolling up different medical products that contain the same active ingredient, and 2) combining records that overlap in time, subject to a persistence window. The first objective is accomplished by leveraging the hierarchy within the standardized terminology to aggregate drugs to the ingredient level of RxNorm. The second objective is achieved by deriving end dates for each drug exposure record, then evaluating whether exposure windows for the same product are sufficiently close to infer continuous use.

For claims related to pharmacy prescriptions, the dispensed date and number of days supply are used to extrapolate the end date for the period of drug exposure. This approach is commonly used and shown to reliably reflect utilization patterns[218]. When a person receives recurring prescriptions for the same product and strength, the multiple prescriptions may need to be treated as a single drug era. To determine whether this is indeed the case, the drug's "persistence window," which is the number of days after the person stops taking a drug and during which the person is deemed to still be affected by the drug, must be taken into account. If the number of days between the end date of the prior Drug Exposure and the start date of the subsequent drug exposure falls within the persistence window, then the two exposures are considered to belong to the same drug era. The 'persistence window' for this experiment is defined to be 30 days.

For example, as illustrated in Figure 15, consider a person who is taking two drugs: Drug A and Drug B. The person has had four prescriptions for Drug A (A1, A2, A3, A4), each with a sixty-day supply. The Person has also had two prescriptions for Drug B (B1, B2).

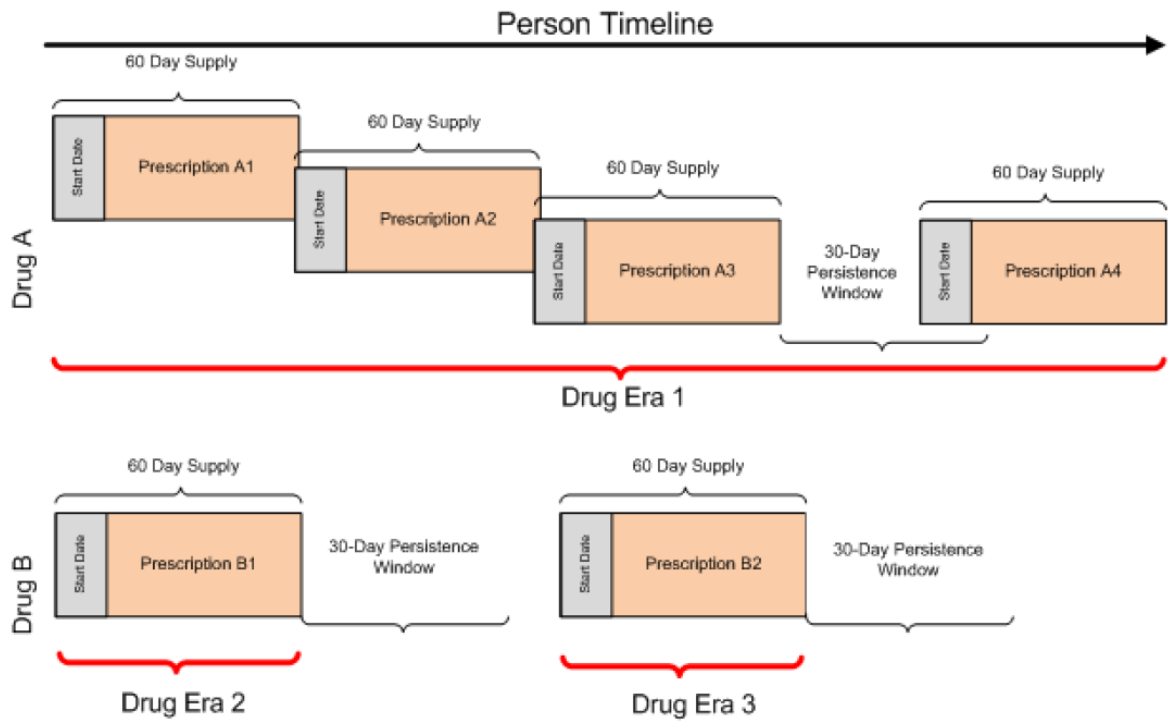


**Figure 15: Drug era construction**

To define the drug era for Drug A, the timing, duration, overlap, and persistence of the person's prescriptions for drug A must be considered. A2 was filled before the expected completion of A1. Similarly, A3 was filled before the expected completion of A2. A4 was filled after A3 was completed, but within the persistence window for Drug A. Therefore, the four prescriptions for Drug A will be consolidated into a single drug era (DrugEra1), with the start for prescription A1 recorded as the start date for the consolidated record and the end date for prescription A4 recorded as the end date. As the persistence window was exceeded between filling the two prescriptions for Drug B, they are defined as two distinct Drug Eras. The start and end dates for DrugEra2 and DrugEra3 are the start and end dates for prescriptions B1 and B2, respectively.

Note, the logic for drug eras does not append overlapping exposure time to the end of the drug exposure length. That is, if a person receives a second 30-day prescription 10 days before the allotted 30-days supply for the first prescription, the resulting drug era would be 50 days long. These ten days will not be added to the persistence window as 'carry over'. This assumes the old prescription was completed or will be used in the future at the time of dispensing or record of the next prescription. Because drugs are rolled up to the ingredient level, this avoids misclassification of dose changes. It could be argued this conservative assumption be revised to augment the exposure length by this overlap, but these assumptions may likely vary by treatment and specific analysis.

In a manner analogous to the construction of drug eras, condition occurrence records are standardized into a common terminology and aggregated into condition eras prior to analysis. Specifically, administrative claims databases code diagnoses as International Classification of Diseases (ICD9) diagnostic codes, while clinical systems use ICD9, Systematized Nomenclature of Medicine-Clinical Terms (SNOMED), and Medical Dictionary for Regulatory Activities (MedDRA). These source codes are mapped into SNOMED as the standard terminology for this analysis.

The common data model contains a CONDITION_OCCURRENCE table, which stores all verbatim records from the source database that could be potentially used to infer condition occurrences. Most source databases provide an identifier for the condition (such as ICD-9-CM diagnosis code) used and a diagnosis date. However, particularly in administrative claims systems, diagnoses may be recorded to facilitate reimbursement of a particular procedure, and may be recorded multiple times on the same or successive dates if more than one service is provided. The CONDITION_ERA table is intended to provide one common structure for aggregating distinct diagnosis records into episodes of care for a given condition.

Similar to Drug Eras, Condition Eras are chronological periods of Condition Occurrence. Combining individual Condition Occurrences into a single Condition Era serves at least two purposes: 1) it allows aggregation of chronic conditions that require frequent ongoing care, instead of treating each Condition Occurrence as an independent event; and 2) it allows aggregation of multiple, closely-timed doctor visits for the same condition to avoid double-counting the Condition Occurrences.

For example, consider a Person who visits his Primary Care Physician (PCP), who diagnoses the Person with a specific condition and refers the Person to a Specialist. One week later, the Person visits the Specialist, who confirms the PCP's diagnosis and provides the appropriate treatment to resolve the condition with no further care required. These two independent doctor visits should be aggregated into one Condition Era.

This model generally fits well for acute conditions, but may be less robust for chronic conditions. For example, chronic conditions that do not require regular follow-up may be recorded as multiple Condition Eras because the absence of data in the periods between visits does not justify the aggregation of the eras. Because the persistence window is small, it is likely that multiple visits will be captured in rapid succession for the same condition; however, it is unlikely that infrequent visits for chronic conditions (e.g. a Person with Rheumatoid Arthritis who visits his rheumatologist every three months) will be captured. However, the small window also reduces the likelihood that independent events will be falsely classified as the same Condition Era.

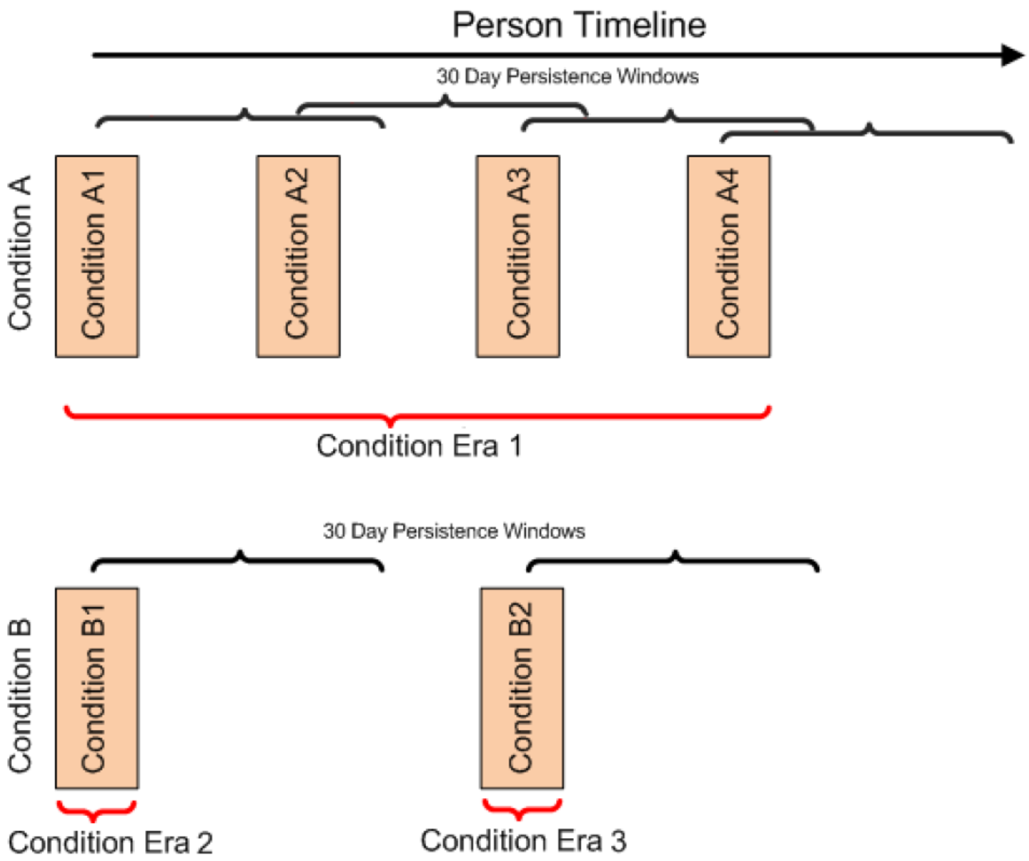


**Figure 16: Condition era construction**

Figure 16 provides an illustration of how the logic for condition eras is applied to diagnosis codes. Imagine a Person who has been diagnosed with two conditions during his insurance coverage period: Condition A and Condition B. The Person has been diagnosed with Condition A four times (A1, A2, A3, A4), and has been diagnosed with Condition B twice (B1, B2).

To define condition persistence for Condition A, the timing of successive diagnoses is considered. Here, A2 is within the persistence window of A1. Similarly, A3 is within the persistence window of A2, and A4 is within the persistence window of A3. Thus, the four diagnoses of Condition A should be consolidated into Condition Era1, with the start date equal to the diagnosis date for A1, and the end date equal to the diagnosis date for A4. With

Condition B, significant time has elapsed between diagnoses B1 and B2. Therefore, it cannot be assumed that there is dependence between the diagnoses as the time exceeded the persistence window for B1. Therefore two distinct Condition Eras are defined, one each that corresponds to B1 and B2.

Note, that for Eras built using 30 day-persistence windows no additional 30 days is being added at the end of the last Condition Occurrence. That means, that Condition-free times within an Era is treated as continual Condition, while in the time following the Era no Condition is assumed. For outcome ascertainment, the condition era onset, and not the era length, is of most direct relevance.

The potential concern with applying any persistence window when defining episodes of care is misclassification. A longer persistence window risks treating diagnoses that reflected independent conditions are part of the same continuum of care, while shorter persistence window assumptions may falsely separate the records from the same episode of care and observe them as distinct occurrences. In the context of active surveillance, where condition occurrences may be used as proxies for potential observations of adverse events, both forms of misclassification bias require careful consideration. Even when using a 30d persistence window assumption, the large majority of aggregated eras come from the same diagnosis occurring less than 10 days from one another. In these cases, it seems more unlikely these conditions represent independent events than it does that the gaps coincide with a common episode of care. A sensitivity analysis of the condition era persistence window was conducted to assess the degree of consolidation at 0 days and 30 days[219]. This analysis shows that between 33% and 45% of records were successfully aggregated in claims databases using 30 day window, and that >70% of gaps between successive diagnoses were within 30 days for the 10 outcomes under study within OMOP.

If a method only uses the first occurrence of cases as a proxy for incident events, then consolidation of eras does not matter (since both have the same first start date). However, if a method attempts to use prevalent cases, as measured by each distinct era occurrence, the selection of the persistence window can be significant. To reiterate, multiple eras for the same condition does not necessarily indicate distinct occurrences of the condition, but instead represent independent periods of time where the data suggests the condition may have occurred. That is, chronic conditions, such as diabetes, are likely to be considered to persist following the first occurrence, but a person may have multiple eras for diabetes because they do not receive care of the disease on a regular basis.

### 3.4 Experimental design

The primary objective of this study is to evaluate the performance of COMPASS as a potential hypothesis generating tool for active drug safety surveillance across a network of observational healthcare databases. Performance is measured as the accuracy by which COMPASS identifies true effects and discerns from false positive findings. The challenge in prospectively evaluating a hypothesis generating tool for drug safety is that 'ground truth' about the drug-outcome relationship is not established; that is, the intended goal of such a tool would be to uncover new safety issues that have not previously been detected. However, new safety issues that are detected may be either true positives or false positive findings, and substantial work in formal evaluation would need to be conducted to confirm or refute any findings. Prior to prospective use of a hypothesis generating tool, it is important to first retrospectively evaluate the performance of the method in an experimental setting so we can establish some level of expectation for prospective performance.

The retrospective evaluation of COMPASS is conducted within six observational databases, and across the drugs within the Angiotensin Converting Enzyme (ACE) Inhibitor class. COMPASS is applied to a series of drug-condition pairs to produce estimates of the potential effects. These estimates are then compared to a pre-defined ‘ground truth’ classification of drug-condition relationships as either positive or negative controls. Measures of accuracy are compared to other hypothesis generating methods to provide both absolute and relative assessment.

ACE Inhibitors provide a solid basis for methodological research because the class represents a large set of mature products that are actively used in the broad population. ACE Inhibitors block the conversion of angiotensin I to angiotensin II within the rennin-angiotensin system, which plays a important role in the pathology of hypertension, cardiovascular health, and renal function[22, 220]. Effective blood pressure reduction has been shown to reduce death, stroke, and heart disease[221]. ACE inhibitors have been found to be effective in the control of hypertension, as well as reduce the risk of acute myocardial infarction among patients with heart failure, left ventricular remodeling after acute myocardial infarction, mortality among patients with severe heart failure and reduced left ventricular ejection fraction, and progression of renal disease among diabetic and non-diabetic patients[220]. Angiotensin II receptor blockers (ARBs) were developed as an alternative treatment option to ACE inhibitors and have been found to have comparable impact on hypertension, cardiovascular disease and heart failure, as well as renal disease progression[220, 222]. Several head-to-head clinical trials and systematic reviews have shown that products with the ACE inhibitor class have comparable efficacy to one another[22, 221], and the class has comparable efficacy to ARBs for the primary indications for both classes[220, 222]. The Joint National Committee on Prevention, Detection, Evaluation and Treatment of High Blood Pressure (JNC-7) currently recommends an ACE inhibitors or ARBs as first line options for patients with stage 1 hypertension who have diabetes, chronic kidney disease, history of stroke or myocardial infarction, or high cardiovascular risk[223]. The American Diabetes Association and Kidney Disease Outcome Quality Initiative guidelines both recommend use of an ACE inhibitors or ARBs for diabetic patients with hypertension or diabetic nephropathy[224], as well as patients with diabetic or non-diabetic proteinuric renal disease[225].

The primary adverse events of ACE inhibitors reported include hypotension, cough, angioedema, hyperkalemia, and acute renal impairment[22]. Other adverse effects include rashes, hepatotoxicity, dysgeusia, and neutropenia. One meta-analysis examined adverse events in 51 placebo- or standard treatment controlled randomized trials of ACE inhibitors in patients with heart failure or ventricular dysfunction, and found that cough (relative risk=1.86), hypotension (RR=1.95), renal dysfunction (RR=1.84), dizziness (RR=1.60), hyperkalemia (RR=7.11), and impotence (RR=6.46) were all significantly more prevalent among patients treated with ACE inhibitors than among those in the control groups[226]. A systematic review comparing ACE inhibitors and ARBs found differences in rate of cough, but no difference in rates of other adverse events such as headache and dizziness[222]. The relative risk of cough was 2.7 in East Asian patients, as compared to whites[227]. When ARBs were added to ACE inhibitor therapy for heart failure, increased risk of hypotension, renal function, and hyperkalemia has been observed[228]. None of the studies have shown significant differences in the rates of cough, angioedema, hyperkalemia, or acute renal impairment between specific ACE inhibitors[22].

While rare in incidence, angioedema has been consistently shown as a potential risk across all ACE inhibitors in clinical trials, and reinforced by observational database studies[70]. Enalapril was shown to have a 4-fold increase in angioedema risk relative to placebo, from 1 per 1,000 to 4 per 1,000 among all subjects[229]. The ALLHAT study demonstrated the same incidence and relative effects in lisinopril, with a rate was 4 per 1,000 for lisinopril users, versus <1 per 1,000 for the other treatments[230]. The HOPE trial showed comparable findings for ramipril[231]. Rates in angioedema were also consistent in trials for captopril[232] and perindopril[233]. The risk of ACE inhibitor-related angioedema is increased in patients of African descent, with an observed two[230] to four-fold[234] increased risk relative to white Americans. The AASK trial showed the significantly different rates of angioedema among ramipril users over 3.5 to 6 years of followup (6.4%), versus 2.3% and 2.7% for metoprolol and amlodipine, respectively[22]. In ALLHAT, rates of angioedema were higher in blacks than non-blacks (0.7% vs 0.3%)[230].

Hypotension (either postural or not defined) was the most consistently reported adverse event was hypotension, but definitions of 'significant' hypotension varied widely between studies, and observed rates varied accordingly[22]. Rates of hypotension among captopril trials ranged from 8% to 37%[22]. Hypotension rates were comparable between ACE inhibitor products, including captopril[232], enalapril[235]**,** and perindopril[233]**.** Hyperkalemia was also consistently reported, and while rates varied significantly in the literature for enalapril[235, 236], no evidence of consistent differences between products in the class. Clinical trials and observational studies have reported renal dysfunction for captopril[232], lisinopril[237], and perindopril[233] with no significant disparities. Two observational studies[232, 233] reported hematological effects, including leucopenia and thrombocytopenia, but did not observe differential effects between drugs. In summary, the ACE inhibitor class has a well-established safety profile, with little evidence to suggest differential effects between products within the class[22].

This study specifically focused on seven medical products within the class: lisinopril, moexipril, quinapril, ramipril, benazepril, captopril, and enalapril. Each ingredient was identified by its corresponding RxNorm ingredient concept identifier, and all clinical drugs (including doses, formulations, and combination products for which an ACE inhibitor is one of the active ingredients) are subsumed through the standard vocabulary relationships. The number of patients exposed to any ACE inhibitor and each of the individual ingredients for each of the six databases is shown in Table 2. Within CCAE, there are over 3 million patients with at least one exposure to an ACE inhibitor, and at least 15,000 patients exposed to each product. Restriction to incident use yield over 1 million persons overall, and greater than 10,000 patients for all ingredients except perindopril. The total sample size varies across the network of databases, but CCAE reflects the largest database and subsequently the largest sample of ACE inhibitor users. However, the proportion of ACE inhibitor users in the Medicare and GE populations are markedly higher than the privately-insured population reflected in CCAE.

Table 3 highlights the FDA-approved indications and off-label uses for each product, as identified using the COMPASS automated heuristic. All products share hypertension as a primary indication, with eight conditions listed as an indication and 14 other conditions being listed as an off-label use for at least one product. However, there are some disparities in secondary indications and off-label uses between products. No two ACE inhibitors share the same indication and off-label use profile. In particular, benazepril is the second-most frequently used ACE inhibitor (behind lisinopril) in all sources, but has the fewest indications reported. The indications are used to construct a comparator group defined by alternative treatments with the same indications, and all indications and off-label uses serve as covariates in the COMPASS propensity model.

Table 4 lists the alternative treatments that share at least one common indication with each product, and therefore would be selected as part of the COMPASS comparator definition. There are 78 ingredients used as a comparator for at least one ACE inhibitor, including products from multiple classes such as Angiotensin II receptor blockers, beta blockers, calcium channel blockers, and diuretics. The 'any ACE inhibitor' analysis constructed a comparator group based on 73 products; the fewest products used in comparator selection is 62 (benazepril) while the largest number of comparator drugs is 74 (ramipril). Note, certain comparator drugs are listed for one ACE inhibitor and not others, due either to differing indications or because the comparator drug may be excluded due to a combination formulation. For example, hydrochlorothiazide is excluded as a comparator for all ACE inhibitors except ramipril because combination products exist, and amlodipine is used for only four ACE inhibitors due to combination use. It is important to note that COMPASS is applying an automated heuristic to comparator selection, and as such, some of the active comparators selected may be different from those that would be defined through expert subjective assessment. For example, some may argue that aspirin and clopidogrel may be inappropriate comparator choices for lisinopril, however they are used for lisinopril due to shared indication of myocardial infarction prevention.

## Table 2: ACE inhibitor use across databases

| | CCAE | MSLR | MDCD | MDCR | GE |
|---|---|---|---|---|---|
| | N=59,836,290 | N=1,466,617 | N=11,188,360 | N=4,655,736 | N=11,216,208 |
| | n (%) | n (%) | n (%) | n (%) | n (%) |
| Any ACE Inhibitor | | | | | |
| Prevalent users | 3,052,264 (5.10) | 108,869 (7.42) | 614,703 (5.49) | 1,569,765 (33.72) | 1,361,058 (12.13) |
| Incident users | 1,137,211 (1.90) | 32,532 (2.22) | 188,224 (1.68) | 483,853 (10.39) | 529,767 (4.72) |
| Lisinopril | | | | | |
| Prevalent users | 1,808,825 (3.02) | 59,039 (4.03) | 374,919 (3.35) | 931,871 (20.02) | 888,890 (7.93) |
| Incident users | 837,280 (1.40) | 21,669 (1.48) | 165,749 (1.48) | 395,600 (8.50) | 399,047 (3.56) |
| Benazepril | | | | | |
| Prevalent users | 576,123 (0.96) | 20,969 (1.43) | 98,911 (0.88) | 253,275 (5.44) | 166,383 (1.48) |
| Incident users | 215,215 (0.36) | 7,228 (0.49) | 38,272 (0.34) | 81,463 (1.75) | 68,539 (0.61) |
| Enalapril | | | | | |
| Prevalent users | 236,215 (0.39) | 12,067 (0.82) | 85,350 (0.76) | 167,866 (3.61) | 130,920 (1.17) |
| Incident users | 85,833 (0.14) | 4,292 (0.29) | 31,800 (0.28) | 51,545 (1.11) | 48,202 (0.43) |
| Ramipril | | | | | |
| Prevalent users | 318,274 (0.53) | 12,356 (0.84) | 76,815 (0.69) | 169,027 (3.63) | 141,059 (1.26) |
| Incident users | 128,343 (0.21) | 3,697 (0.25) | 15,441 (0.14) | 67,553 (1.45) | 65,580 (0.58) |
| Quinapril | | | | | |
| Prevalent users | 195,047 (0.33) | 7,684 (0.52) | 36,146 (0.32) | 95,355 (2.05) | 88,094 (0.79) |
| Incident users | 37,571 (0.06) | 1,476 (0.10) | 9,498 (0.08) | 15,341 (0.33) | 33,933 (0.30) |
| Captopril | | | | | |
| Prevalent users | 43,613 (0.07) | 1,685 (0.11) | 23,822 (0.21) | 67,609 (1.45) | 31,854 (0.28) |
| Incident users | 11,360 (0.02) | 345 (0.02) | 6,618 (0.06) | 13,195 (0.28) | 10,761 (0.10) |
| Moexipril | | | | | |
| Prevalent users | 43,152 (0.07) | 840 (0.06) | 7,253 (0.06) | 23,262 (0.50) | 17,856 (0.16) |
| Incident users | 11,501 (0.02) | 164 (0.01) | 1,805 (0.02) | 4,928 (0.11) | 8,063 (0.07) |

## Table 3: Indication covariates identified by COMPASS for each ACE inhibitor

| Indication | ACE Inhibitors | Lisinopril | Benazepril | Enalapril | Ramipril | Quinapril | Captopril | Moexipril |
|---|---|---|---|---|---|---|---|---|
| Asymptomatic Left Ventricular Dysfunction | A | | | A | | | | |
| Bartter's Syndrome | O | | | | | | O | |
| Chronic Heart Failure | A | A | O | A | A | A | A | O |
| Cystine Renal Calculi | O | | | | | | O | |
| Cystinuria | O | | | | | | O | |
| Diabetic Nephropathy | A | O | O | O | O | O | A | O |
| Diabetic Retinopathy | O | O | | | | | | |
| Diagnostic Test for Primary Aldosteronism | O | | | | | | O | |
| Diastolic Heart Failure | O | O | O | O | O | O | O | O |
| Edema | O | | | | | | O | |
| Hypertension | A | A | A | A | A | A | A | A |
| Hypertension due to Scleroderma | O | | O | O | O | | O | |
| Hypertensive Emergencies | O | | | | | | O | |
| Left Ventricular Dysfunction following Myocardial Infarction | A | | | | A | | A | |
| Migraine Prevention | O | O | | | | | | |
| Myocardial Infarction | A | A | | | | | | |
| Myocardial Infarction Prevention | A | | | | A | | | |
| Nondiabetic Proteinuric Nephropathy | O | O | O | O | O | O | O | O |
| Prevention of Cerebrovascular Accident | A | | | | A | | | |
| Prevention of Recurrent Atrial Fibrillation | O | O | O | O | O | O | O | O |
| Raynaud's Phenomenon | O | | | | | | O | |
| Renal Crisis Scleroderma | O | O | O | O | O | O | O | |

A: FDA-approved indication; O-off-label use

## Table 4: Comparator drugs selected by COMPASS for each ACE inhibitor

| Drug class | Comparator drug | ACE Inhibitors | Lisinopril | Benazepril | Enalapril | Ramipril | Quinapril | Captopril | Moexipril |
|---|---|---|---|---|---|---|---|---|---|
| Aldosterone Receptor Antagonists | Spironolactone | X | X | X | X | X | X | X | X |
| Aldosterone Receptor Antagonists | eplerenone | X | X | X | X | X | X | X | X |
| Alpha-Beta Blockers | Labetalol | X | X | X | X | X | X | X | X |
| Alpha-Beta Blockers | carvedilol | X | X | X | X | X | X | X | X |
| Analgesic - Central Alpha-2 Receptor Agonists | Clonidine | X | X | X | X | X | X | X | X |
| Angiotensin II Receptor Blockers (ARBs) | Losartan | X | X | X | X | X | X | X | X |
| Angiotensin II Receptor Blockers (ARBs) | Olmesartan medoxomil | X | X | X | X | X | X | X | X |
| Angiotensin II Receptor Blockers (ARBs) | candesartan | X | X | X | X | X | X | X | X |
| Angiotensin II Receptor Blockers (ARBs) | eprosartan | X | X | X | X | X | X | X | X |
| Angiotensin II Receptor Blockers (ARBs) | irbesartan | X | X | X | X | X | X | X | X |
| Angiotensin II Receptor Blockers (ARBs) | telmisartan | X | X | X | X | X | X | X | X |
| Angiotensin II Receptor Blockers (ARBs) | valsartan | X | X | X | X | X | X | X | X |
| Antianginal - Coronary Vasodilators (Nitrates) | Nitroglycerin | X |  |  |  | X |  | X |  |
| Antiarrhythmic - Class IV | Diltiazem |  | X |  |  | X | X | X | X |
| Antiarrhythmic - Class IV | Verapamil |  | X |  |  | X | X | X | X |
| Anticoagulants - Coumarin | Dicumarol | X | X |  |  |  |  |  |  |
| Antihyperlipidemic - HMG CoA Reductase Inhibitors (statins) | Lovastatin | X |  |  |  | X |  |  |  |
| Antihyperlipidemic - HMG CoA Reductase Inhibitors (statins) | Pravastatin | X |  |  |  | X |  |  |  |
| Antihyperlipidemic - HMG CoA Reductase Inhibitors (statins) | Simvastatin | X |  |  |  | X |  |  |  |
| Antihyperlipidemic - HMG CoA Reductase Inhibitors (statins) | atorvastatin | X |  |  |  | X |  |  |  |
| Beta Blockers Cardiac Selective | Atenolol | X | X | X | X | X | X | X | X |
| Beta Blockers Cardiac Selective | Betaxolol | X | X | X | X | X | X | X | X |
| Beta Blockers Cardiac Selective | Bisoprolol | X | X | X | X | X | X | X | X |
| Beta Blockers Cardiac Selective | Metoprolol | X | X | X | X | X | X | X | X |
| Beta Blockers Cardiac Selective | nebivolol | X | X | X | X | X | X | X | X |
| Beta Blockers Cardiac Selective, Intrinsic Sympathomimetic Activity | Acebutolol | X | X | X | X | X | X | X | X |
| Beta Blockers Non-Cardiac Select., Intrinsic Sympathomimetic Activity | Carteolol | X | X | X | X | X | X | X | X |
| Beta Blockers Non-Cardiac Select., Intrinsic Sympathomimetic Activity | Penbutolol | X | X | X | X | X | X | X | X |

| Drug class | Comparator drug | ACE Inhibitors | Lisinopril | Benazepril | Enalapril | Ramipril | Quinapril | Captopril | Moexipril |
|---|---|---|---|---|---|---|---|---|---|
| Beta Blockers Non-Cardiac Select., Intrinsic Sympathomimetic Activity | Pindolol | X | X | X | X | X | X | X | X |
| Beta Blockers Non-Cardiac Selective | Nadolol | X | X | X | X | X | X | X | X |
| Beta Blockers Non-Cardiac Selective | Propranolol | X | X | X | X | X | X | X | X |
| Beta Blockers Non-Cardiac Selective | Timolol | X | X | X | X | X | X | X | X |
| Calcium Channel Blockers - Dihydropyridines | Amlodipine |  | X |  |  | X | X | X | X |
| Calcium Channel Blockers - Dihydropyridines | Felodipine |  | X |  |  | X | X | X | X |
| Calcium Channel Blockers - Dihydropyridines | Isradipine | X | X | X | X | X | X | X | X |
| Calcium Channel Blockers - Dihydropyridines | Nicardipine | X | X | X | X | X | X | X | X |
| Calcium Channel Blockers - Dihydropyridines | Nifedipine | X | X | X | X | X | X | X | X |
| Calcium Channel Blockers - Dihydropyridines | Nisoldipine | X | X | X | X | X | X | X | X |
| Calcium Channel Blockers - Dihydropyridines | clevidipine | X | X | X | X | X | X | X | X |
| Calcium Channel Blockers - T-Type Channel Acting Agents | Mibefradil | X | X | X | X | X | X | X | X |
| Central Alpha-2 Receptor Agonists | Guanabenz | X | X | X | X | X | X | X | X |
| Central Alpha-2 Receptor Agonists | Guanfacine | X | X | X | X | X | X | X | X |
| Central Alpha-2 Receptor Agonists | Methyldopa | X | X | X | X | X | X | X | X |
| Central Alpha-2 Receptor Agonists | Methyldopate | X | X | X | X | X | X | X | X |
| Digitalis Glycosides | Digoxin | X | X |  | X | X | X | X |  |
| Direct Acting Vasodilators | Hydralazine | X | X | X | X | X | X | X | X |
| Direct Acting Vasodilators | Minoxidil | X | X | X | X | X | X | X | X |
| Direct Acting Vasodilators | Nitroprusside | X | X | X | X | X | X | X | X |
| Diuretic - Loop | Furosemide | X | X | X | X | X | X | X | X |
| Diuretic - Loop | torsemide | X | X | X | X | X | X | X | X |
| Diuretic - Potassium Sparing | Amiloride | X | X | X | X | X | X | X | X |
| Diuretic - Thiazides and Related | Bendroflumethiazide | X | X | X | X | X | X | X | X |
| Diuretic - Thiazides and Related | Chlorothiazide | X | X | X | X | X | X | X | X |
| Diuretic - Thiazides and Related | Chlorthalidone | X | X | X | X | X | X | X | X |
| Diuretic - Thiazides and Related | Hydrochlorothiazide |  |  |  |  | X |  |  |  |
| Diuretic - Thiazides and Related | Hydroflumethiazide | X | X | X | X | X | X | X | X |
| Diuretic - Thiazides and Related | Indapamide | X | X | X | X | X | X | X | X |
| Diuretic - Thiazides and Related | Methyclothiazide | X | X | X | X | X | X | X | X |
| Diuretic - Thiazides and Related | Metolazone | X | X | X | X | X | X | X | X |
| Diuretic - Thiazides and Related | Polythiazide | X | X | X | X | X | X | X | X |

| Drug class | Comparator drug | ACE Inhibitors | Lisinopril | Benazepril | Enalapril | Ramipril | Quinapril | Captopril | Moexipril |
|---|---|---|---|---|---|---|---|---|---|
| Diuretic - Thiazides and Related | Trichlormethiazide | X | X | X | X | X | X | X | X |
| Ganglionic Blocking, Non-Depolarizing | Mecamylamine | X | X | X | X | X | X | X | X |
| Peripheral Alpha-1 Receptor Blockers | Doxazosin | X | X | X | X | X | X | X | X |
| Peripheral Alpha-1 Receptor Blockers | Prazosin | X | X | X | X | X | X | X | X |
| Peripheral Alpha-1 Receptor Blockers | Terazosin | X | X | X | X | X | X | X | X |
| Platelet Aggregation Inhibitors - Salicylates | Aspirin | X | X |  |  | X |  |  |  |
| Platelet Aggregation Inhibitors - Thienopyridine Agents | clopidogrel | X | X |  |  |  |  |  |  |
| Postganglionic Blockers, Antihypertensive | Guanethidine | X | X | X | X | X | X | X | X |
| Postganglionic Blockers, Antihypertensive | guanadrel | X | X | X | X | X | X | X | X |
| Renin Inhibitor, Direct | aliskiren | X | X | X | X | X | X | X | X |
| Reserpine and Derivatives | Reserpine | X | X | X | X | X | X | X | X |
| Reserpine-Thiazide & Related Combinations | benzothiazide | X | X | X | X | X | X | X | X |
| Thrombolytic - Tissue Plasminogen Activators | Alteplase | X | X |  |  |  |  |  |  |
| Unclassifed | Alseroxylon | X | X | X | X | X | X | X | X |
| Unclassifed | Pargyline | X | X | X | X | X | X | X | X |
| Unclassifed | Phenprocoumon | X | X |  |  |  |  |  |  |
| Unclassifed | cyclothiazide | X | X | X | X | X | X | X | X |
| Unclassifed | quinethazone | X | X | X | X | X | X | X | X |

X: drug used in comparator group

**Table 5: Contraindications used as restriction criteria by COMPASS for each ACE inhibitor X: condition listed as contraindication**

| Indication | ACE Inhibitors | Lisinopril | Benazepril | Enalapril | Ramipril | Quinapril | Captopril | Moexipril |
|---|---|---|---|---|---|---|---|---|
| Abnormal Hepatic Function Tests | X | X | X |  |  |  |  |  |
| Acute Hepatic Failure | X | X |  |  |  |  |  |  |
| Acute Pancreatitis | X | X | X | X |  | X | X | X |
| Acutely Decompensated Chronic Heart Failure | X |  |  | X |  |  |  |  |
| Anaphylaxis occuring from Desensitization to Allergens | X |  |  |  | X |  |  |  |
| Angioedema | X | X | X | X |  | X | X | X |
| Anuria | X | X | X | X |  | X | X | X |
| Aortic Valve Stenosis | X | X | X | X |  |  | X | X |
| Ascites | X |  |  |  | X |  |  |  |
| Atrial Fibrillation with Lown-Ganong-Levine Syndrome | X |  |  | X |  |  |  |  |
| Atrial Fibrillation with Wolff-Parkinson-White | X |  |  | X |  |  |  |  |
| Azotemia | X |  |  |  | X |  |  |  |
| Bone Marrow Depression | X | X | X | X |  | X | X | X |
| Bradycardia | X |  |  | X |  |  |  |  |
| Cardiogenic Shock | X |  |  | X |  |  |  |  |
| Cerebrovascular Insufficiency | X | X |  | X |  | X | X | X |
| Chronic Heart Failure | X |  |  | X |  |  |  |  |
| Chronic Idiopathic Constipation | X |  | X | X |  |  |  |  |
| Complete Atrioventricular Block | X |  |  | X |  |  |  |  |
| Connective Tissue Disease | X | X | X | X | X | X | X | X |
| Cough | X |  |  |  |  |  | X |  |
| Dehydration | X |  | X |  | X | X |  |  |
| Diabetes Mellitus | X | X | X | X |  | X | X | X |
| Disease of Liver | X | X | X | X | X | X | X | X |
| Disorder of Electrolytes | X | X |  |  |  |  |  |  |
| Gout | X | X | X | X |  | X | X | X |
| Head and Neck Angioedema | X |  | X |  | X |  |  |  |
| Hemodialysis with High-Flux Membrane | X | X | X | X | X | X | X | X |

| Indication | ACE Inhibitors | Lisinopril | Benazepril | Enalapril | Ramipril | Quinapril | Captopril | Moexipril |
|---|---|---|---|---|---|---|---|---|
| Hepatic Cirrhosis | X |  |  |  | X |  |  |  |
| Hepatic Coma | X | X | X | X |  | X | X | X |
| Hereditary Angioedema | X | X | X | X |  | X | X | X |
| Hypercalcemia | X | X | X | X |  | X | X | X |
| Hypercholesterolemia | X | X | X | X |  | X | X | X |
| Hyperkalemia | X | X | X | X | X | X | X | X |
| Hyperparathyroidism | X | X | X | X |  | X | X | X |
| Hypertrophic Cardiomyopathy | X | X |  | X |  |  | X | X |
| Hyperuricemia | X | X | X | X |  | X | X | X |
| Hypokalemia | X | X | X | X |  | X | X | X |
| Hypomagnesemia | X | X | X | X |  | X | X | X |
| Hyponatremia | X | X | X | X | X | X | X | X |
| Hypotension | X | X | X | X | X | X | X | X |
| Hypovolemia | X | X |  |  |  |  |  |  |
| Immunosuppression | X | X |  | X |  | X | X | X |
| Incomplete AV Heart Block | X |  |  | X |  |  |  |  |
| Intestinal Angioedema | X | X | X |  | X |  |  |  |
| Jaundice | X | X | X |  |  |  |  |  |
| Left Ventricular Dysfunction following Myocardial Infarction | X |  |  | X |  |  |  |  |
| Myocardial Infarction | X |  |  | X |  |  |  |  |
| Myocardial Ischemia | X |  |  |  |  | X |  |  |
| Neonatal Hyperbilirubinemia | X | X | X | X |  | X | X | X |
| Neutropenic Disorder | X | X | X | X | X | X | X | X |
| Oliguria | X | X | X | X | X | X | X | X |
| Peripheral Edema | X |  |  | X |  |  |  |  |
| Pregnancy | X | X | X | X | X | X | X | X |
| Renal Artery Stenosis | X | X | X | X | X | X | X | X |
| Renal Disease | X | X | X | X | X | X | X | X |
| SIADH Syndrome | X | X |  |  |  |  |  |  |
| Scleroderma | X |  |  |  | X |  |  |  |
| Severe Aortic Valve Stenosis | X |  |  | X |  |  |  |  |
| Severe Coronary Artery Disease | X |  | X | X |  |  |  |  |

| Indication | ACE Inhibitors | Lisinopril | Benazepril | Enalapril | Ramipril | Quinapril | Captopril | Moexipril |
|---|---|---|---|---|---|---|---|---|
| Severe Diarrhea | X | | | | X | | | |
| Severe Hepatic Disease | X | | X | | | | | |
| Severe Hypotension | X | | X | X | | | | |
| Severe Renal Disease | X | X | X | X | | X | X | X |
| Severe Vomiting | X | | | | X | | | |
| Sick Sinus Syndrome | X | | | X | | | | |
| Sympathectomy | X | X | X | X | | X | X | X |
| Systemic Lupus Erythematosus | X | X | X | X | X | X | X | X |
| Transplantation Procedure | X | | | X | | | X | X |
| Ventricular Tachycardia | X | | | X | | | | |

The safety profile of ACE inhibitors is thought to be well-characterized, including a broad set of known safety issues that span the continuum from common, nuisance effects, such as cough, to rare and more series events, like angioedema and renal dysfunction. To conduct the retrospective evaluation of COMPASS, we must first define a reference set of 'positive controls' and 'negative controls' that can be used to assess methods' performance. This objective was achieved through a systematic analysis of the structured product labels, using labeled events as surrogate markers for 'positive controls' and selecting terms unrelated to any labeled events as 'negative controls'.

Regenstrief Institute has developed a novel application, Structured Product Label Information Coder and Extractor (SPLICER), which performs natural language processing on structured product labels (SPL) to extract terms that may be adverse events. SPLs are FDA-approved labeling from product manufacturers, publicly available from the National Library of Medicine, that is formatted in XML to facilitate standardized evaluation. The application classifies the events by the location of occurrence, as 'Black box', 'Warnings and Precautions', 'Adverse Reactions' or 'Post-marketing experience', and codes the terms of MedDRA preferred terms (PTs). Each SPL was mapped to a corresponding RxNorm drug concept.

SPLICER's most recent run was performed on 12/19/2009 and include 5602 SPL's from the DailyMed site[238]. This set of labels comprised 1706 distinct generic drugs and 2861 distinct brand names. SPLICER successfully coded and extracted 608,948 adverse events from these labels. These events were mapped to 4627 distinct MedDRA preferred terms. An evaluation of SPLICER's performance in retrieving events from the Adverse Reaction section of 100 labels demonstrated a recall of 93% and a precision of 95%. The output of SPLICER for SPLs from the ACE inhibitors was used for defining the reference set. Table 6 provides a descriptive summary of the number of events extracted from each SPL, summarized by the ingredients within each DOI. For example, there are 28 distinct SPLs that involve lisinopril. Among those 28 labels, SPLICER identified 234 distinct MedDRA PTs. On average, each of the labels listed 184 distinct events, with the minimum of 60 events and a maximum of 205 events. This table highlights the variability observed in product labeling, both among labels for the same ingredient as well as among ingredient within the same drug class.

**Table 6: Labeled events identified in SPLs by ingredient**

| Ingredient name | Number of SPLs | Distinct events across SPLs | Min events among SPLs | Average events among SPLs | Max events among SPLs |
|---|---|---|---|---|---|
| Lisinopril | 28 | 234 | 60 | 184 | 205 |
| Moexipril | 6 | 261 | 72 | 158 | 242 |
| Quinapril | 15 | 174 | 72 | 101 | 151 |
| Ramipril | 9 | 112 | 87 | 100 | 110 |
| Benazepril | 19 | 180 | 58 | 89 | 133 |
| Captopril | 12 | 143 | 103 | 114 | 135 |
| Enalapril | 15 | 211 | 117 | 142 | 171 |
| Fosinopril | 12 | 183 | 116 | 132 | 140 |
| Perindopril | 3 | 153 | 150 | 151 | 152 |

### Selecting 'positive controls'

Labeled events were selected as 'positive controls' test cases if three criteria were satisfied:

1. MedDRA PT was listed on >=50% of structured product labels within the OMOP DOI

2. MedDRA PT had at least one ICD-9-CM code directly mapped to it within the OMOP standardized terminology
3. One of the ICD-9-CM codes mapped to the MedDRA PT also directly mapped to at least one SNOMED concept

We created two levels of classification: Tier 1 events are those conditions that occur in either the 'Black box' or 'Warnings/Precautions' sections on >=50% of the SPLs within the class. Tier 2 events are those conditions that occur as adverse events anywhere on the product label (Black box, warnings/precautions, adverse reactions, or Post-marketing experience) on >=50% of the SPLs within the class. Tier 1 events are a subset of the Tier 2 labeled events. It could be argued that events listed in black box warnings or warnings/precautions are more likely to be causally related and observable. Primary analyses will be based on all Tier 2 events, but Tier 1 classification offers a potential sensitivity analysis when assessing methods performance.

The rationale for criteria #1 was that the majority of labels contributing to the drug class needed to list the event in order to have some confidence that the association could be potentially observed. The rationale for criteria #2 and #3 is to ensure the event is theoretically observable across the data sources under study. That is, some MedDRA preferred terms include adverse event concepts that have no corresponding codes in ICD-9-CM, so could not possibly be recorded in any US administrative claims system. Another issue is that some ICD-9 codes may map to multiple concepts; in these cases, the ICD-9 code is mapped in the standardized terminology to a surrogate concept and is excluded from consideration as a test case. We chose to restrict our focus to concepts that also have corresponding SNOMED concepts to enable us to evaluate both MedDRA and SNOMED as alternative standardized terminologies for active surveillance.

As a class, ACE Inhibitors have 84 SNOMED-based ‘true positives’, 21 of which are Tier 1 Warning events. The list of the terms is shown in Table 7. The observed background prevalence in CCAE is categorized for each condition.

**Table 7: ACE Inhibitor 'true positive' reference set**

| ConceptID | Condition Concept Name | Prevalence | Position in Label |
|---|---|---|---|
| 196490 | Acute renal failure following labor AND/OR delivery | Low | Tier 1 Warning |
| 197320 | Acute renal failure syndrome | High | Tier 1 Warning |
| 316447 | Chronic hypotension | High | Tier 1 Warning |
| 254761 | Cough | High | Tier 1 Warning |
| 22350 | Edema of larynx | High | Tier 1 Warning |
| 193782 | End stage renal disease | High | Tier 1 Warning |
| 197988 | Generalized abdominal pain | High | Tier 1 Warning |
| 316866 | Hypertensive disorder | High | Tier 1 Warning |
| 193519 | Impaired renal function disorder | High | Tier 1 Warning |
| 317002 | Low blood pressure | High | Tier 1 Warning |
| 314432 | Maternal hypotension syndrome | Low | Tier 1 Warning |
| 313829 | Maternal hypotension syndrome - delivered with postnatal problem | Medium | Tier 1 Warning |
| 31967 | Nausea | High | Tier 1 Warning |
| 27674 | Nausea and vomiting | High | Tier 1 Warning |
| 75365 | Oliguria and anuria | High | Tier 1 Warning |
| 196764 | Post-delivery acute renal failure - delivered with postnatal problem | Low | Tier 1 Warning |
| 4167493 | Pregnancy-induced hypertension | Medium | Tier 1 Warning |
| 195014 | Renal failure following molar AND/OR ectopic pregnancy | Medium | Tier 1 Warning |
| 4058979 | Renal sclerosis NOS | None | Tier 1 Warning |
| 433879 | Umbilical pain | High | Tier 1 Warning |
| 441408 | Vomiting | High | Tier 1 Warning |
| 440979 | Acquired hemolytic anemia | High | Tier 2 Label |
| 440372 | Acquired thrombocytopenia | High | Tier 2 Label |
| 4110022 | Acute bronchitis and bronchiolitis | Medium | Tier 2 Label |
| 23798 | Acute laryngopharyngitis | High | Tier 2 Label |
| 258453 | Acute upper respiratory infection of multiple sites | High | Tier 2 Label |
| 139902 | Allergic urticaria | High | Tier 2 Label |
| 439777 | Anemia | High | Tier 2 Label |
| 321318 | Angina | High | Tier 2 Label |
| 73231 | Arthralgia of temporomandibular joint | High | Tier 2 Label |

| ConceptID | Condition Concept Name | Prevalence | Position in Label |
|---|---|---|---|
| 78508 | Arthralgia of the ankle and/or foot | High | Tier 2 Label |
| 77642 | Arthralgia of the forearm | High | Tier 2 Label |
| 81112 | Arthralgia of the lower leg | High | Tier 2 Label |
| 79106 | Arthralgia of the pelvic region and thigh | High | Tier 2 Label |
| 78516 | Arthralgia of the upper arm | High | Tier 2 Label |
| 437113 | Asthenia | High | Tier 2 Label |
| 317009 | Asthma | High | Tier 2 Label |
| 438727 | Atypical depressive disorder | Medium | Tier 2 Label |
| 256717 | Bronchospasm | High | Tier 2 Label |
| 77670 | Chest pain | High | Tier 2 Label |
| 256448 | Chronic asthmatic bronchitis | High | Tier 2 Label |
| 4110029 | Chronic pharyngitis and nasopharyngitis NOS | Medium | Tier 2 Label |
| 75860 | Constipation | High | Tier 2 Label |
| 136775 | Contact dermatitis due to solar radiation | High | Tier 2 Label |
| 75635 | Cramp | High | Tier 2 Label |
| 196523 | Diarrhea | High | Tier 2 Label |
| 312437 | Dyspnea | High | Tier 2 Label |
| 433440 | Dysthymia | High | Tier 2 Label |
| 318556 | Epistaxis | High | Tier 2 Label |
| 437448 | Exhaustion due to excessive exertion | Medium | Tier 2 Label |
| 436297 | Exhaustion due to exposure | Medium | Tier 2 Label |
| 318566 | Flushing | High | Tier 2 Label |
| 80141 | Functional diarrhea | High | Tier 2 Label |
| 440674 | Gout | High | Tier 2 Label |
| 440071 | Gout associated problem | Medium | Tier 2 Label |
| 78234 | Hand joint pain | High | Tier 2 Label |
| 23325 | Heartburn | High | Tier 2 Label |
| 194087 | Hepatitis due to infection | Medium | Tier 2 Label |
| 138565 | Hyperhydrosis disorder | High | Tier 2 Label |
| 434610 | Hyperkalemia | High | Tier 2 Label |
| 312950 | IgE-mediated allergic asthma | High | Tier 2 Label |
| 436962 | Insomnia | High | Tier 2 Label |
| 252658 | Intrinsic asthma without status asthmaticus | High | Tier 2 Label |
| 77074 | Joint pain | High | Tier 2 Label |
| 435224 | Leukopenia | High | Tier 2 Label |
| 194133 | Low back pain | High | Tier 2 Label |
| 78517 | Multiple joint pain | High | Tier 2 Label |
| 437834 | Non-autoimmune hemolytic anemia | Medium | Tier 2 Label |
| 375838 | Objective tinnitus | Medium | Tier 2 Label |
| 314666 | Old myocardial infarction | High | Tier 2 Label |
| 319041 | Orthostatic hypotension | High | Tier 2 Label |
| 315078 | Palpitations | High | Tier 2 Label |
| 135338 | Pemphigus | Medium | Tier 2 Label |
| 78162 | Peripheral vertigo | High | Tier 2 Label |
| 134159 | Precordial pain | High | Tier 2 Label |
| 441264 | Primary thrombocytopenia | High | Tier 2 Label |
| 4067066 | Pruritus and related conditions | Low | Tier 2 Label |

| ConceptID | Condition Concept Name | Prevalence | Position in Label |
|---|---|---|---|
| 136184 | Pruritus of skin | High | Tier 2 Label |
| 441540 | Reactive confusion | High | Tier 2 Label |
| 78232 | Shoulder joint pain | High | Tier 2 Label |
| 140821 | Spasm | High | Tier 2 Label |
| 381864 | Subjective tinnitus | High | Tier 2 Label |
| 377575 | Tinnitus | High | Tier 2 Label |
| 4181583 | Upper respiratory infection | High | Tier 2 Label |

### Selecting 'negative controls':

Labeled events were selected as 'negative control' test cases if four criteria were satisfied:

1. MedDRA PT does not have the same High Level Term as any PT that was extracted from any location (black box, warnings/precautions, adverse reactions, post-marketing experience, indications) of any structured product labels among any drug
2. MedDRA PT had at least one ICD-9-CM code directly mapped to it within the OMOP standardized terminology
3. One of the ICD-9-CM codes mapped to the MedDRA PT also directly mapped to at least one SNOMED concept
4. MedDRA PT belongs to a System Organ Class other than "Pregnancy, puerperium and perinatal conditions" and "Congenital, familial and genetic disorders"

Criteria #1 ensures that no 'negative control' is related to any labeled event. This is a conservative restriction to avoid selecting any terms that could be drug-related by eliminating all adverse events that occurred on at least one label. The 'negative control' must exist in a High Level Term without any other labeled events to minimize the chance that a 'negative control' would be selected because it was a distinct term even though it was clinically similar.  For example, ‘myocardial infarction’ is a labeled event, but ‘acute myocardial infarction’ and ‘myocardial ischemia’ are not; however, since all three terms belong to the HTL ‘Coronary ischemic disorders’,  all are excluded as candidate ‘negative controls’. Criteria #4 was applied because pregnancy-related adverse events are often ill-defined and typically reflect special case circumstances that are not the specific focus of this study.

Based on these criteria, 2,800 distinct SNOMED terms were identified as ‘negative controls’.

**Table 8: ACE Inhibitor negative controls, by prevalence**

| Prevalence | SNOMED terms |
|---|---:|
| High | 608 |
| Medium | 1439 |
| Low | 724 |
| None | 29 |
| Total | 2800 |

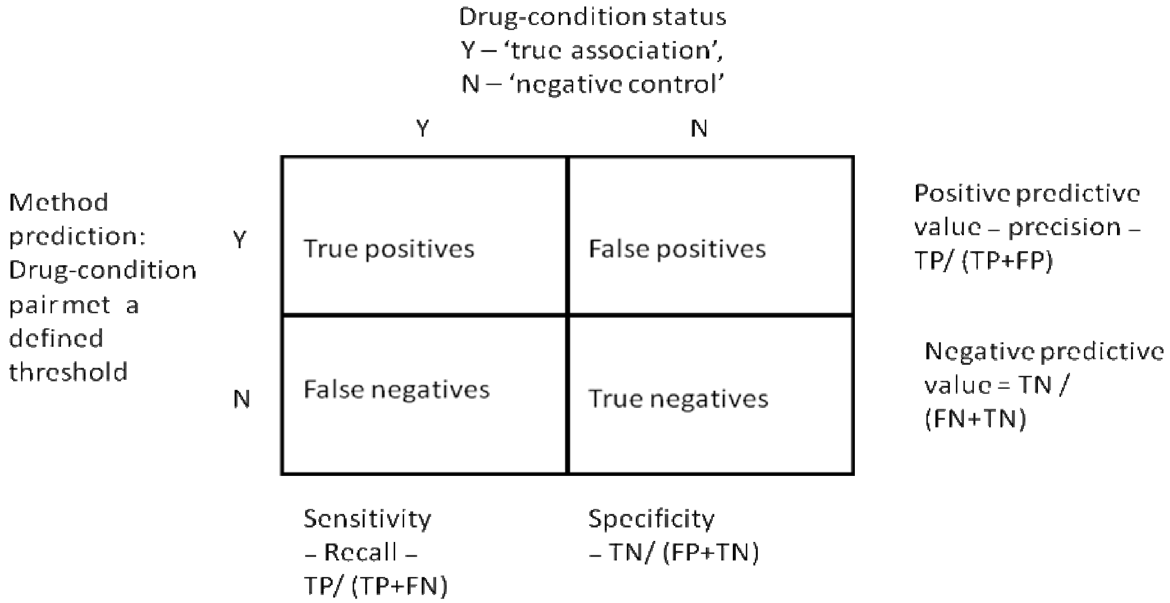
It is acknowledged that the objective heuristic used to construct the reference set, both true positives and negative controls, is subject to misclassification.  Because ‘truth’ is not known for any drug, we are required to select some surrogate (which has its own undefined sensitivity and specificity).  We understand that labeled events have not necessarily been shown to be causally related to drug, or may not be expected to be observed in subsequent study.  In particular, adverse events listed in the Adverse Reactions and Post-Marketing Experience section may reflect occurrence from clinical trials or spontaneous reporting without any expectation of causality.  Similarly, it is possible that ‘negative controls’ have been selected that do have legitimate temporal relationships with the drugs of interest, and either have not been previously identified or were not listed on the product label.   For purposes of the experiment, all scores for ‘negative controls’ that suggest a relationship were classified as false positive findings.  The process used to identify the reference set was empirically driven to minimize subjective assessment, but carries its own limitations. SPLICER may misclassify adverse events, either missing or failing to code events that exist on the label or identifying terms on the labels that are not actual adverse events. While the application has strong performance characteristics for the Adverse Reactions section, it may be more prone to error in the Black Box or Warnings/Precautions sections due to the unstructured nature of the text. SPLICER classifies all matched terms meeting its criteria as potential adverse events, though may misclassify terms that were instead risk factors or contraindications.

That said, it is not necessary to identify all potential ‘positive controls’ or all eligible ‘negative controls’. Instead, the number of test cases can be considered the sample size within this methodological experiment. Because the same set of test cases is being consistently applied across all methods, any misclassification of test cases (either ‘positive controls’ that are not related, or ‘negative controls’ that have an association) should not introduce differential bias to the experiment and should not influence the relative assessment of performance measures between methods.

A key limitation in this experimental design is the potential lack of generalizability in the results. A method’s performance in identifying known drug safety issues and discerning from known non-issues may not be consistent with performance of classifying unknown effects. Because the performance characteristics calculated are based on the artificial definition of truth used for experiment, care should be taken when attempting to predict how methods may perform prospectively in an active surveillance network. Instead, these metrics should be considered to be most appropriate for comparative purposes across methods and databases. The experiment’s use of six data sources should provide a robust measure of performance across disparate data, but the findings may not be directly applicable to a network of different data sources that could potentially be used for a national active surveillance system. While ACE inhibitors offer a wide array of drug safety issues to test again, the results may not be generalizable to all potential effects expected to be detected within an active surveillance system. In particular, the performance against other medical products, such as newly marketed medicines with low initial use, products for other therapeutic uses, and drugs with acute or intermittent exposure, may vary from results observed in this study.

### 3.5 Performance measures

The potential use of COMPASS as a hypothesis generating tool for identifying drug safety issues is analogous to signal detection theory, and measures of performance that follow from diagnostic and screening testing are well suited for study. The aim is to predict a binary classification of drug-condition status (there is, or is not, a causal relationship between exposure and outcome). The method prediction is a continuous valued score, but could be imagined to be dichotomized at some defined threshold. In this context, the test cases could be categorized into the following 2x2 contingency table (Figure 17), and various measures of performance can be estimated.



Drug-condition status
Y – 'true association',
N – 'negative control'

| Method prediction: Drug-condition pair met a defined threshold | | Y | N | |
|---|---|---|---|---|
| | Y | True positives | False positives | Positive predictive value – precision – TP/ (TP+FP) |
| | N | False negatives | True negatives | Negative predictive value = TN / (FN+TN) |
| | | Sensitivity – Recall – TP/ (TP+FN) | Specificity – TN/ (FP+TN) | |

**Figure 17: Performance measures for 2x2 contingency table**

Measures of accuracy can be applied within the experiment that are not constrained to defined dichotomization of the method score. In addition to studying COMPASS performance at logical thresholds, such as ARDLB>0, the performance of COMPASS was characterized through multiple measures of accuracy, including mean average precision, precision-at-k, and area under receiver operator characteristic (ROC) curve.

'Mean average precision' (MAP) can be thought of as the average precision at each threshold value that represents a 'true positive' association. MAP is effectively the equivalent to the area under precision-recall curve. MAP can be formally defined as follows. Let $y_{dc} = 1$ if the *d*th drug is associated with the *c*th condition ('positive control') and zero otherwise, *d*=1,…,*D*, *c*=1…,*C*. Let $M = \sum_{d,c} y_{dc}$ denote the number of causal combinations and $N = D \times C$ the total number of combinations. Let $z_{dc}$ denote the predicted value for the *d*th drug and the *c*th condition. For a given set of predicted values $\vec{z} = (z_{11}, \ldots, z_{dc})$, we define "precision-at-*K*" denoted $P^{(K)}(\vec{z})$ as the fraction of causal combinations amongst the *K* largest predicted values in $\overset{\rho}{z}$ . Specifically, let $z_{(1)} > \Lambda > z_{(N)}$ denote the ordered values of $\overset{\rho}{z}$ . Then:

$$P^{(K)}(\overset{\rho}{z}) = \frac{1}{K}\sum_{i=1}^{K} y_{(i)} \text{ ,}$$

where $y_{(i)}$ is the true status of the combination corresponding to $z_{(i)}$. "Mean Average Precision" is then defined as:

$$S = \frac{1}{M}\sum_{K:y_{(K)}=1} P^{(K)}(\overset{\rho}{z})$$

Unscored conditions are treated as if they produced a minimum score, such that methods receive the maximum penalty for not classifying 'positive controls'.

'Precision-at-k' (P@k) is commonly used in information retrieval, and reflects the proportion of correctly classified objects at a defined cutoff (k) among an ordered set. So, in drug safety contexts, setting k=100, P@k could be interpreted as: 'among the top 100 estimates produced by the method, what proportion of the drug-condition pairs reflect positive controls'.

An additional tool for assessing accuracy is the Receiver Operator Characteristic (ROC) curve, which are based on evaluating true positive rate (sensitivity) and false positive rate (1-specificity). The area under the ROC curve (AUC) provides a scalar measure of performance at all potential thresholds.

Finally, we define 'recall-at-FP' (R@fp) as the sensitivity obtained at a defined tolerance of false positive rate. So, for example, setting FP=5%, R@fp can be interpreted as:

‘what proportion of true positives can a method identify before 5% of negative controls would also be identified’.

Mean average precision, precision-at-k, area under curve, and recall-at-fp all provide scalar measures of performance, but each reflect a complementary component for interpretation. None are sufficient, since each have inherent limitations. Precision-at-k and recall-at-fp are inherently threshold-based, insofar as a subjective assessment of k and fp is required. In contrast, MAP and AUC are threshold-independent, but provide a composite score that may reflect boundary conditions of little practical use. For example, AUC integrates over all levels of specificity, including high false positive rates that would likely be unacceptable in a drug safety context. Similarly, MAP integrates over all levels of recall, though it may be unrealistic to expect that a given method can identify all adverse events with high precision and focus on more modest levels of detection may be more appropriate. A method that produces higher performance scores across all summary measures can be considered to have superior aggregate performance. However, it is feasible for methods to have differential behavior across the summary measures.

Moreover, summary performance measures do not reflect expectations for performance for any specific adverse event, as each condition can have different attributes (such as background prevalence, time-to-onset, strength of association, and degree of confounding) that could alter a method’s behavior for that relationship. For each drug-condition pair, a method produces a score, but the performance of that pair cannot be measured without putting the score into context with other scores produced by the method for other drug-condition pairs. As such, for each event, it is possible to measure precision and false positive rate at the score produced by essentially treating the event score as the threshold for dichotomizing scores, as shown in Figure 17. Event-based performance measures were provided to explore differential method performance across the positive controls.

### 3.5 Data analysis by Aim

**Aim 1: Characterize the performance of COMPASS in identifying known safety issues associated with ACE inhibitor exposure within an administrative claims database**

This aim studied how COMPASS performs in the Thomson Reuters MarketScan Commercial Claims and Encounters (CCAE), a large administrative claims database containing 59 million privately insured lives. COMPASS was applied to the ACE Inhibitor drug class to generate estimates of outcome relationships for a defined set of 2884 potential adverse events. These outcomes include both the 84 known associations previously characterized in the product label as well as a sample of 2800 'negative control' conditions for which there is no evidence of drug-related effects. Each test case reflects a condition concept in the SNOMED terminology that subsumes one or more ICD9 codes.

Descriptive statistics summarized the distribution of the estimates and patterns across attributes of the conditions, such as the ground truth status, background prevalence rate, confidence in association, and expected degree of confounding. Stratified probability density functions were used to explore two-way interactions between ground truth and condition characteristics.

The objective of a hypothesis-generating tool is to accurately distinguish between true and false relationships. The performance of COMPASS were characterized through multiple measures of accuracy, including mean average precision, area under ROC, precision-at-k, and recall-at-fp. These measures provide an estimate of overall performance across all test cases. In addition, we evaluated the tradeoff between four performance characteristics (sensitivity, specificity, positive predictive value, and negative predictive value) at alternative threshold values, including ARD LBCI > 0, which would be a natural indicator for designating a significant relationship. For each ‘true positive’, we characterized performance by assessing the false positive rate and precision if the threshold were set at the test case score. This review across specific conditions allows the exploration of differential performance among the true relationships.

These measures were compared to those from three alternative methods for active surveillance signal generation: disproportionality analysis; observational screening; and, univariate self-controlled case series. Each method has previously been proposed for use in active surveillance, yet use fundamentally different analytical strategies for producing drug-condition estimates. Disproportionality analysis reflects an adaption to a data mining signal detection approach used in spontaneous adverse event reporting. Observational screening applies an unadjusted cohort-based design to compare event rates during exposure to that of the overall population. Self-controlled case series is a case-based approach that attempts to measure drug effects based on time exposed and unexposed among those with at least one outcome. All three approaches have been made publicly available as part of the OMOP methods library, and are described in further detail below. For each alternative method, the same descriptive statistics used for COMPASS were applied to facilitate comparisons. In addition, because method scores are measured on different scales and may have different degrees of variability, the rank of scores from each method were  used to enable visualization of relative performance across methods on a normalized scale.

The comparison of these hypothesis generating tools is exploratory in nature.  As such, there is no formal statistical test being applied to determine that COMPASS is superior or non-inferior to other alternative approaches.  Such tests do exist for comparing AUC between alternative diagnostic tests.  However, in active surveillance, there is an expectation that no single method will be sufficient for all potential active surveillance needs, and that multiple approaches may be useful across a network of disparate data sources.  This is due to the heterogeneity that exists within the potential drug-condition associations under potential study.  Unlike a diagnostic tool for a defined condition, such as DXA for detecting hip fracture, where the variability in the tool's performance is inherent to the individual being assessed, the variability in an active surveillance tool stems from variability of the adverse event, the exposure, and the source population.  For example, overall performance as measured by MAP or AUC may suggest the use of a particular tool, but for a specific condition with a particular set of characteristics, an alternative approach may be preferred.  Instead, the objective of this study is to determine if COMPASS has the potential to complement existing approaches.

### Disproportionality analysis

Disproportionality analysis methods for drug safety surveillance represent the primary class of analytic methods for analyzing data from spontaneous adverse event reporting systems (SRSs). SRSs receive reports that comprise of one or more drugs, one or more adverse events (AEs), and possibly some basic demographic information (in addition to narrative and text data). Disproportionality analysis methods include the multi-item gamma-Poisson shrinker (MGPS), proportional reporting ratios (PRR), reporting odds ratios (ROR), and Bayesian confidence propagation neural network (BCPNN). The methods search SRS databases for "interesting" associations and focus on low-dimensional projections of the data, specifically 2-dimensional contingency tables, as shown below.

| | AE j = Yes | AE j = No | Total |
|---|---|---|---|
| Drug i = Yes | $w_{00}$ | $w_{01}$ | $w_{0*}$ |
| Drug i = No | $w_{10}$ | $w_{11}$ | $W_{1*}$ |
| Total | $w_{*0}$ | $w_{*1}$ | $w_{**}$ |

Given a two-by-two table such as Table 2, various disproportionality metrics can be estimated as shown below.

Proportional reporting ratio[154]:

$$PRR = \frac{w_{00}/w_{00} + w_{01}}{w_{10}/w_{10} + w_{11}}$$

Reporting odds ratio[155]:

$$ROR = \frac{w_{00}/w_{10}}{w_{01}/w_{11}}$$

Multi-item Gamma Poisson Shrinker[9]:

Let $w_{00}(i,j)$ denote the $w_{00}$ entry for the two-by-two table for the $i$th drug and the $j$th condition. Assume that each $w_{00}(i,j)$ is a draw from a Poisson distribution with mean $m(i,j)$. Let $m(i,j) = l(i,j)*\mathrm{E}(i,j)$, where $\mathrm{E}(i,j)=w_{0+}(i,j)*w_{+1}(i,j)/w_{++}(i,j)$, i.e., the expected value of $w_0(i,j)$ under independence and is assumed to be known. The goal is to estimate the values of the $l$'s . A $l(i,j)$ far from one supports the notion that drug $i$ and condition $j$ are not independent. MGPS is a Bayesian procedure and starts with a particular five-parameter prior distribution for the collection of *l*'s:

$$\pi(\lambda;\alpha_1,\beta_1,\alpha_2,\beta_2,P) = Pg(\lambda;\alpha_1,\beta_1) + (1-P)g(\lambda;\alpha_2,\beta_2)$$

where $g(\lambda;\alpha,\beta)$ denotes a gamma density with *a*/*b*. The "EBGM" measure is defined as:

$$EBGM(i,j) = 2^{EB\log_2(i,j)}$$

where:

$$EB\log_2 = \Big(Q_w\big[\psi(\alpha_1 + w_{00} - \log(\beta_1 + E)\big] + (1-Q_w)\big[\psi(\alpha_n + w_{00} - \log(\beta_2 + E)\big]\Big)/\log(2)$$

$$Q_w = Pf(w_{00};\alpha_1,\beta_1,E)/\big[Pf(w_{00};\alpha_1,\beta_1,E) + (1-P)f(w_{00};\alpha_2,\beta_2,E)\big] \text{ and}$$

$$f(w_{00};\alpha,\beta,E) = (1+\beta/E)^{-w_{00}}(1+E/\beta)^{-\alpha}\,\Gamma(\alpha+w_{00})/\Gamma(\alpha)n!.$$

MGPS uses an empirical Bayes approach and chooses $a_1$, $b_1$, $a_2$, $b_2$, and *P* to maximize:

$$\prod_{i,j} Pf(w_{00}(i,j);\alpha_1,\beta_1,E(i,j)) + (1-P)f(w_{00}(i,j);\alpha_2,\beta_2,E(i,j)).$$

The EBGM score is the mean of the posterior distribution of the true RR. Other summaries are possible. For example, DuMouchel mentions "EB05". This is the 5th percentile of the posterior distribution – meaning that there is a 95% probability that the "true" RR exceeds the EB05. Since EB05 is always smaller than EBGM this, in a sense, adds extra shrinkage and represents a more conservative choice than EBGM.

Bayesian confidence propagation neural network (BCPNN) Information Component (IC)[165]:

$$IC(i,j) = \log_2 \frac{w_{00}(i,j) + 1/2}{E(i,j) + 1/2}$$

Disproportionality analysis methods can be readily applied to longitudinal observational databases insofar as the longitudinal data can also be projected into 2x2 contingency tables. Various design alternatives are available for defining this projection. These decisions include: how to count events (whether to categorize distinct patient status or replicate the notion of spontaneous reports of conditions following exposure); definition of outcomes based on incident or prevalent occurrence of conditions; definition of a surveillance window to infer time-at-risk relative to exposure start or end (30 days from onset; all time exposed, time exposed + 30 days from end, any time following exposure start); and whether to stratify on age, gender and/or year of report to calculate expected values. Combined with the various potential metrics, there are 112 configurations of the OMOP disproportionality analysis under experimentation.

### Observational screening

Observational screening is a method originally developed at GlaxoSmithKline and now made commercially available as part of the SAEfetyWorks® software application by ProSanos[18, 193-196]. Screening applies a basic cohort design to estimate the rate of condition occurrence during the time exposed to a particular product d:

$$SR_d = \frac{\sum_i x_i}{\sum_i t_i}$$

where for the i-th person, $x_i$ is the number of conditions that occurred during the time-at-risk $t_i$, as defined by the periods of exposure (drug era end date – drug era start date) and some surveillance window.

This screening rate is then compared to the overall background rate of the condition to produce a screening rate ratio:

$$SRR_d = SR_d / \frac{\sum_i X_i}{\sum_i T_i}$$

where $SR_d$ is the screening rate for the drug of interest d, and $X_i$ is the number of conditions that occurred any time within the i-th person's observation period (regardless of exposure status) and $T_i$ is the total observation time.

The estimate is unadjusted and therefore susceptible to various forms of confounding. The primary intent, in its original conception, was that screening rate ratios could be calculated very efficiently for all potential drug-condition pairs across large observational databases, and could provide a first-pass approach for identifying potential issues that warrant further evaluation. The screening rate ratio metric is a crude estimate of the absolute effect size, which could be used to identify differences in occurrence of outcomes during exposure. Screening assumes the screening rate ratio is a ratio of two Poisson distributed rates, and uses the closed form solution by Graham et al to estimate confidence interval:

UB95: $(t_1/t_0)*((2*x_1*x_0+Z_{\alpha/2}{}^2*(x_1+x_0) + \surd( Z_{\alpha/2}{}^2*(x_1+x_0)*(4*x_1*x_0+ Z_{\alpha/2}{}^2*(x_1+x_0))))/(2*(x_1)^2))$

LB95: $(t_1/t_0)*((2*x_1*x_0+Z_{\alpha/2}{}^2*(x_1+x_0) - \surd( Z_{\alpha/2}{}^2*(x_1+x_0)*(4*x_1*x_0+ Z_{\alpha/2}{}^2*(x_1+x_0))))/(2*(x_1)^2))$

where $t_0$ is person-time exposure for cohort 0,

$t_1$ is person-time exposure for the entire data source,

$x_0$ is the number of events occurring during exposure in cohort 0,

$x_1$ is the number of events occurring at any time for the entire data source

Various design decisions can alter the screening rate estimate, some of which are under experimentation within the OMOP implementation of observational screening. These include: the use of first exposure or all exposures; defining outcomes based on incident conditions, prevalent conditions, or first condition appearing within time-at-risk; definition of a surveillance window to infer time-at-risk relative to exposure start or end (30 days from onset; all time exposed, time exposed + 30 days from end, any time following exposure start); and whether to count as outcomes conditions that occur on the first day of exposure. Additionally, both the screening rate ratio point estimate and lower bound of the confidence interval can be used as scores to use for prioritizing potential effects. In total, 32 different parameter settings are explored as potential alternative configurations of observational screening.

**Self-controlled case series**

The univariate self-controlled case series (USCCS) approach assumes that events arise among persons as a non-homogeneous Poisson process[19, 20, 239]. The method only makes use of persons who have time exposed and unexposed, and also have experienced at least one event. The observation period for person i is the time period during which an event could be observed. Each person's observation period is split into risk periods, indexed by j. Let $e_{ij}$ denote the time spent by individual i in risk period j. The incidence, denoted $\lambda_{ij}$, is assumed to be constant within each interval. The current implementation of univariate self-controlled case series assumes a multiplicative model for the incidence function: $\lambda_{ij} = \exp(\Phi_i + \beta_j)$ where $\Phi_i$ represents an effect for each person i, and $\beta_j$ represents an effect for risk group j, with $\beta_0 = 0$. The incidence function during the baseline period is simply $\lambda_{i0} = \exp(\Phi_i)$. Note, other logical extensions can be applied, such as further risk modeling based on age groups, concomitant drug use, or other time-varying covariates[20], but are not included in the current model proposed for active surveillance. Conditioning on the number of events $n_i$ observed for person i during the observation period, the log likelihood is multinomial:

$$l(\beta) = \sum_{ij} n_{ij} log \left[ \frac{\exp(\beta_j) e_{ij}}{\sum_s \exp(\beta_s) e_{is}} \right]$$

The desirable phenomenon within the self-controlled case series framework is that all person-level effects $\Phi_i$ cancel out, because incidence rates are compared within a given person's time window. In the active surveillance context, USCCS can be applied across multiple drugs and conditions, but the estimates of the drug-condition relationships are treated independently. The estimate β can be used as a relative measure of effect, and is produced as the score for each drug-condition pair by the OMOP self-controlled case series program. Within the self-controlled case series framework, several design decisions are required that are under experimentation within OMOP. These include: whether to define outcomes based on incident or prevalent occurrence of conditions; whether to include the first day of exposure in the time-at-risk; definition of a surveillance window to infer time-at-risk relative to exposure start or end (30 days from onset; all time exposed, time exposed + 30 days from end, exposure + 60 days from end); and, precision of the Normal prior (0.5, 0.8, 1, 2). Measures of standard error for the univariate design can be estimated, but are not included within the current approach under consideration. All combinations of potential parameter settings are empirically evaluated to produce 64 distinct self-controlled case series analyses.

### Aim 2: Evaluate consistency of COMPASS estimates across a network of disparate databases

An active surveillance network is likely to comprise multiple data sources, as it is recognized that there is currently no single US-based source that can be expected to satisfy all requirements of allowing investigation of all medical products for all potential adverse events and across all populations of interest. However, there is little research to inform the expected behavior of active surveillance analysis methods when applied to disparate databases, or the potential benefits of integrating estimates across sources to improve method performance.

This aim conducted the COMPASS analysis for ACE inhibitors across five databases. Beyond CCAE, the method was applied to the MarketScan Lab Database (MSLR), MarketScan Medicaid Multi-State Database (MDCD), MarketScan Medicare Supplemental and Coordination of Benefits Database (MDCR), and the GE Centricity electronic health record (GE). For each database, COMPASS was applied to the ACE Inhibitor drug class to generate estimates of outcome relationships for the same set of 2884 test cases (84 'positive controls' and 2800 'negative controls').

Four accuracy measures (mean average precision, area under ROC, precision-at-k, and recall-at-fp) were calculated for each database to measure overall performance across the test cases. In addition, for each 'true positive', we characterized the performance in each source by assessing the false positive rate and precision at the test case score. These accuracy measures from each source were compared to assess the reliability in performance.

$I^2$ statistics were calculated to assess the heterogeneity in COMPASS estimates across data sources. Scatterplots were used to explore the relationships among scores between each of the 21 pairwise combinations of data sources.

In addition, we explored three approaches to producing composite estimates based on the individual scores provided from each contributing data source. The first composite will apply a simple threshold heuristic to categorize test cases based on the number of sources that produce a statistically significant estimate: $C_1 = \sum_i (ARD\ LBCI_i > 0)$, where i is the index for the sources. Here, the composite score measures how many sources corroborate the association, with the expectation that greater sources with significant relationships reflect increased confidence in a potential relationship (though each source contributes equally to the measure). Note, when $C_1 = 5$, then in effect, this approach provides a conservative assessment that requires all sources to corroborate an association before the condition is considered a potential issue.

The first composite score is proposed because it provides a simple heuristic that has been suggested as a potential approach to consider within a distributed network of active surveillance systems. A more formal approach is to pool the estimates and measures of uncertainty within a meta-analytic framework. The composite score $C_2$ is based on a pooled rate difference from fixed effect model using the inverse variance method[46, 240]:

$$C_2 = \bar{RD} = \frac{\sum_{i=1}^{k} \left[ w_i \left( \frac{a_i}{t_{ei}} - \frac{b_i}{t_{ci}} \right) \right]}{\sum_{i=1}^{k} w_i}$$

$$w_i = \frac{{t_{ei}}^2 * {t_{ci}}^2}{a_i {t_{ei}}^2 - b_i {t_{ci}}^2}$$

$$SE(RD) = \frac{1}{\sqrt{\sum w_i}}$$

$$RD\ CI = \bar{RD} \pm z_{1-a/2} * SE(\bar{RD})$$

Where i is the data source, k is the total number of sources, $a_i$ is the number of events in the exposed group, $t_{ei}$ is the person-time in the exposed group, $b_i$ is the number of events in the comparator group, $t_{ci}$ is the person-time in the comparator group, and $w_i$ reflects the inverse of the estimated variance for each study.

The heterogeneity statistic is given by:

$$Q = \sum w_i (RD_i - \overline{RD})^2$$

$I^2$ = 100% * (Q – d.f.) / Q

Q follows a chi-square distribution with k-1 degrees of freedom under the null hypothesis that the true treatment effect is the same for all sources. $I^2$ for the homogeneity test will be provided for each outcome[241]. A distribution of heterogeneity measures was provided to determine the degree to which source variability influences estimation across the range of true positives and negative controls. Given the diversity in the data sources, we fully expect the potential for significant heterogeneity between sources. In fact, the heterogeneity may be sufficiently large that the use of meta-analysis pooled estimates could be questioned. Certainly, in the contexts of producing a valid estimate for a formal evaluation, considerations around the sources of variability and how they may influence effect estimates require specific attention. However, in the context of active surveillance, where we are trying to produce estimates to generate hypotheses about potential effects, we are primarily concerned the degree to which a composite estimate provides a more reliable screening tool, even in spite of observed heterogeneity.

We also explored a DerSimonian and Laird random effects model to relax the assumption that there is a common treatment effect across the sources[46]. Here, we assume the true effect follows a Normal distribution with mean and variance $\tau^2$, where:

$$\tau^2 = \frac{Q - (k-1)}{\sum w_i - \left(\frac{\sum w_i^2}{\sum w_i}\right)}$$

$$w_i{}' = \frac{1}{SE(RD)^2 + \tau^2}$$

$$C_3 = \overset{\frown}{\bar{R}\bar{D}} = \frac{\sum_{i=1}^{k}\left[w_i{}'\left(\frac{a_i}{t_{ei}} - \frac{b_i}{t_{ci}}\right)\right]}{\sum_{i=1}^{k} w_i{}'}$$

$$SE(RD) = \frac{1}{\sqrt{\sum w_i{}'}}$$

It seemed initially compelling to consider using a random-effects meta-regression technique to further adjust for study-level covariates, such as database demographics (age and gender distribution), data capture characteristics, and potentially the accuracy measure (AUC). However, despite having access to 6 large data sources, we are underpowered for meta-regression because the unit of analysis is the study, not the population, and we do not have sufficient degrees of freedom with 6 estimates to produce a reliable composite summary.

Using these composite estimates, we then used the same measures of performance as described in Aim 1 to evaluate will then evaluate the relative performance of the pooled estimate in predicting drug safety issues as compared to source-specific performance to assess the potential advantages of a network-based approach to active surveillance. It is expected the composite estimates should have improved performance to the source-specific estimates, both due to pooling data to increase power as well as the minimization of source-specific effects that could lead to false positive findings.

### Aim 3: Explore differential effects across ingredients within ACE inhibitor class

The general consensus within the clinical community is that all ACE inhibitors have similar safety profiles. However, examination of the product labels suggests differences in which adverse events have been reported. Further, there is little information to assess the relative effect size of adverse events across products in a real-world setting. This aim applied COMPASS to seven medical products within the class (lisinopril, moexipril, quinapril, ramipril, benazepril, captopril, enalapril), to determine whether meaningful differences are observed within observational databases. In this aim, COMPASS was used as a hypothesis generating tool to highlight potential disparities in adverse event rates between products that may warrant further evaluation. The results of this exploratory analysis should not be considered definitive; indeed, differences will be observed through indirect comparisons of adjusted risk differences. A formal pharmacoepidemiology evaluation would likely design a study that provides a direct assessment of the relative effect, and would tailor the analysis to address the specific adverse event of interest. In this context, as an initial active surveillance tool, COMPASS is used to identify the differences between products to facilitate prioritization of effects that may require this additional analysis.

Table 9 highlights the events that were explored within this study. Among the conditions, 17 events were consistently recorded across the product labels for all nine ingredients. These events include events listed as warnings, such as cough, hypotension, and renal dysfunction. The product labels do not provide evidence about the anticipated effect size, or whether the risks should be anticipated to be differential among specific ingredients. In absence of any additional information, it could be assumed that the effects should be shown to be consistent for these 17 events. COMPASS will be applied to these events to generate hypotheses about potential differential risk profiles.

The additional six events in the table reflect adverse events that are not consistently reported in the product labels. For example, epistaxis and tinnitus are listed on product labels for all ingredients except quinapril and captopril. Ramipril does not list asthma, flushing, low back pain, or bronchospasm as potential effects, although the majority of the other products do include these events. These disparities may reflect true differences in observed adverse events, or could simply reflect artifacts of the product labeling standards and the lack of enforcement of consistency across product manufacturers. COMPASS was applied to these six events to discern whether risk differences can be observed across the products.

**Table 9: Adverse events to explore across ACE inhibitor ingredients**

| SNOMED term | Products with event on label | Lisinopril | benazepril | Enalapril | Ramipril | quinapril | Captopril | moexipril |
|---|---|---|---|---|---|---|---|---|
| Consistent events | | | | | | | | |
| Acquired hemolytic anemia | 9 | X | X | X | X | X | X | X |
| Constipation | 9 | X | X | X | X | X | X | X |
| Cough | 9 | X | X | X | X | X | X | X |
| Diarrhea | 9 | X | X | X | X | X | X | X |
| Dyspnea | 9 | X | X | X | X | X | X | X |
| End stage renal disease | 9 | X | X | X | X | X | X | X |
| Generalized abdominal pain | 9 | X | X | X | X | X | X | X |
| Impaired renal function disorder | 9 | X | X | X | X | X | X | X |
| Leukopenia | 9 | X | X | X | X | X | X | X |
| Low blood pressure | 9 | X | X | X | X | X | X | X |
| Nausea | 9 | X | X | X | X | X | X | X |
| Oliguria and anuria | 9 | X | X | X | X | X | X | X |
| Orthostatic hypotension | 9 | X | X | X | X | X | X | X |
| Palpitations | 9 | X | X | X | X | X | X | X |
| Primary thrombocytopenia | 9 | X | X | X | X | X | X | X |
| Pruritus of skin | 9 | X | X | X | X | X | X | X |
| Vomiting | 9 | X | X | X | X | X | X | X |
| Inconsistent events | | | | | | | | |
| Epistaxis | 7 | X | X | X | X | 🟨 | 🟨 | X |
| Tinnitus | 7 | X | X | X | X | 🟨 | 🟨 | X |
| Asthma | 6 | X | X | X | 🟨 | X | X | X |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Flushing | 6 | X | X | X | 🟨 | 🟨 | X | X |
| Low back pain | 6 | X | X | X | 🟨 | X | 🟨 | X |
| Bronchospasm | 5 | X | 🟨 | X | 🟨 | 🟨 | X | X |

COMPASS produced an adjusted rate difference and associated confidence interval for each ingredient-outcome pair. These estimates within each outcome can then be compared to determine if two products have potentially differential effects.

# CHAPTER FOUR: MANUSCRIPT 1:
# "Systematic identification of drug safety issues in administrative claims data: Performance of hypothesis generation methods for active surveillance"

## Abstract

There is emerging interest to expand the use of observational databases, such as administrative claims and electronic health records, as part of an active surveillance network to identify potential drug safety concerns in a more timely manner. However, few studies have evaluated the operational characteristics of the methods proposed for such surveillance. This study explored the performance of three existing methods (disproportionality analysis, observational screening, and self-controlled case series) and introduced a new approach, Comparator-Adjusted Safety Surveillance (COMPASS), which augments an inception cohort design with automated heuristics for comparator selection, inclusion/exclusion criteria, and covariate adjustment through propensity score stratification. Methods were evaluated in a large administrative claims database to assess their ability to identify true safety concerns and discern from false positive findings associated with ACE inhibitor exposure. COMPASS generated the fewest safety signals, had the lowest false positive rate, highest predictive probability and greatest precision of the four methods. Self-controlled case series achieved higher sensitivity but lower specificity. The proposed COMPASS method is a new alternative analysis approach to consider in developing a national active surveillance system, but further methodological research is needed to improve the utility of all methods as hypothesis generating tools.

## Background

Safety assessment of medical products involves a wide array of information. Prior to regulatory approval, pre-clinical toxicology studies, pharmacology experiments and clinical trials provide initial assessment of adverse drug reactions, but are limited both in generalizability to real-world populations and in size for detection of less common events or reactions with modest increased risks from background observed rates (1). In the post-approval setting, spontaneous adverse event reporting offers the opportunity for patients and providers to notify FDA and product manufacturers of adverse experiences post-exposure. However, this passive surveillance system suffers significant limitations for providing a complete safety assessment, including event underreporting, reporting bias, incomplete information and lack of follow-up (2). An additional source of post-approval safety information has been the conduct of analytic pharmacoepidemiologic evaluation studies, which are typically defined to explore a specific hypothesis about a drug-related effect within a real-world population. Observational healthcare databases, such as administrative claims and electronic health records, have provided useful information for pharmacoepidemiologists to conduct these retrospective studies by applying a formal study design to an available dataset in order to estimate the magnitude of the effect between a particular exposure and outcome (3). While pharmacoepidemiologic studies are often less resource-intensive than randomized studies, they require significant expertise and several months before the customized assessment of the individual hypothesis is completed. The often intractable challenges of confounding in observational studies require substantial effort to address and often limit the confidence the community places in these studies, in relation to other available experimental evidence.

The increasing availability of these data sources, coupled with recent information technology innovations, has raised interest in expanding the use of large linked healthcare data to create an active drug safety surveillance system that would complement current practice. The active surveillance system is envisioned to “actively search for patterns in prescription, outpatient and inpatient data systems that might suggest the occurrence of an adverse event, or safety signal, related to drug therapy” (4). Unlike the existing use of pharmacoepidemiologic studies to study pre-defined hypotheses of individual drugs and outcomes at a particular timepoint, the active surveillance system would be applied across a network of disparate databases continuously over time to both generate and refine hypotheses of potential issues associated with all regulated medical products and across a large array of potential adverse events.

In the US, the development of a national active surveillance system is being coordinated by the FDA under the Sentinel Initiative (5), but little evidence is available to inform best practices about appropriate methods to use or expected operating characteristics for such a system once it comes online. The Observational Medical Outcomes Partnership (OMOP) was established to conduct methodological research for the national active surveillance system (6), and has provided a public forum for experimentation amongst data holders and methods developers to begin to address some of these outstanding research questions. Several methods have been proposed as potential approaches for active surveillance, including disproportionality analysis, as adapted from spontaneous data mining(7,8); observational screening, an unadjusted cohort-based design(9); and univariate self-controlled case series(10,11). All of these methods were implemented within the OMOP community (12). In this study, we introduce a new approach, Comparator-Adjusted Safety

Surveillance (COMPASS), which applies a propensity score cohort design using automated heuristics for key design elements, including comparator selection, inclusion/exclusion criteria, and covariate adjustment. The overall aim of this study is to measure the performance of these alternative analysis methods for active surveillance. To address this, we performed a retrospective evaluation of all four methods against a large administrative claims database, assessing each method's performance in their ability to properly classify adverse events with their known association with ACE inhibitors.

## Materials and Methods

### Data

The study population used for the evaluation came from the Thomson Reuters MarketScan Commercial Claims and Encounters (CCAE), a large administrative claims database containing 59 million privately insured lives. CCAE provides patient-level de-identified data from inpatient and outpatient visits and pharmacy claims of multiple large employer-based health plans from 2003 to 2008. CCAE contains 3,052,264 persons with at least one prescription dispensing record for an ACE inhibitor, though each method uses a different fraction of that sample based on a particular study design. The CCAE database was transformed into an OMOP common data model [CDM], with *International Classification of Diseases*, Ninth Revision (ICD-9) diagnosis codes translated into a standardized terminology using condition concepts from *Systematized Nomenclature of Medicine-Clinical Terms* (SNOMED CT).

## Method

### Disproportionality analysis

Disproportionality analysis (DP) methods were developed for use in analyzing spontaneous adverse event reporting databases by identifying drug-event combinations that were co-reported more frequently than what would be expected had the drug and event been independent. Spontaneous reports can be used to construct a series of 2x2 contingency tables, one for each drug-event combination, on the basis of whether the report contains the drug of interest and whether the report contains the event of interest. Several metrics can be applied to these 2x2 tables to produce estimates of the association, including multi-item gamma-Poisson shrinker (MGPS), proportional reporting ratios (PRR), reporting odds ratios (ROR), and Bayesian confidence propagation neural network (BCPNN) (7,8). Disproportionality analysis methods can be readily applied to longitudinal observational databases insofar as the longitudinal data can also be projected into 2x2 contingency tables.

### Observational screening

Observational screening (OS) is a method originally developed at GlaxoSmithKline and now made commercially available as part of the SAEfetyWorks® software application by ProSanos (9, 13,14). Observational screening applies a basic cohort design to estimate the rate of condition occurrence during the time exposed to a particular product, and compares that rate to the overall background rate of the condition in the overall population. The estimate is unadjusted and therefore susceptible to various forms of confounding. The primary intent, in its original conception, was that screening rate ratios could be calculated efficiently for all potential drug-condition pairs across large observational databases, and could provide a first-pass approach for identifying potential issues that warrant further evaluation.

**Self-controlled case series**

The univariate self-controlled case series (USCCS) approach assumes that events arise among persons as a non-homogeneous Poisson process (15). The method only makes use of persons who have are both exposed and unexposed during the observation period, and also have experienced at least one event. The method estimates the relative rate within each person by evaluating the rate of events during the exposed and unexposed time, before producing a composite effect estimates across all cases. The self-controlled design thus is unconfounded by patient characteristics that are stable over time.

**COMPASS**

COMParator-Adjusted Safety Surveillance (COMPASS) is a statistical algorithm that estimates adjusted rate ratios for all outcomes of interest for a given medical product through propensity score stratification across exposed and unexposed cohorts. COMPASS applies an automated heuristic for defining a comparator group based on the indication of the medical product, and provides multivariate adjustment based on key risk factors, including person demographics, comorbidity, and health service utilization. Figure 1 highlights the conceptual model that serves as the basis for COMPASS. The fundamental goal of a drug safety analysis is to assess the temporal relationship between treatment and outcome. However, in the context of an active surveillance system that leverages longitudinal healthcare databases in a non-experimental design, specific attention is needed to minimize bias when estimating the drug-outcome association. COMPASS applies a retrospective cohort design to compare the effects of the target drug of interest to an unexposed population, defined as those exposed to an alternative treatment for the same indication. The COMPASS model focuses on minimizing bias from four primary sources: personal demographics (such as age and gender), confounding by indication, effects of comorbidity, and health service utilization.

The COMPASS approach incorporates several notable features into its analysis that bear particular consideration. First, it incorporates large biomedical ontologies, or networks of clinical concepts such as relationships between diseases and treatments, to automate comparator selection by identifying all drugs that share at least one FDA-approved indication but have different mechanisms of action than the target drug of interest. Second, it imposes automated study design heuristics, including cohort exclusion criteria based on contraindications and covariate selection based on FDA-approved indications and off-label uses. Third, the use of a comorbidity index and multiple measures of health service utilization as additional aggregate covariates allows for improved balancing of exposed and unexposed cohorts that are universally applicable for all outcomes while minimizing concerns of inflating bias due to unconfounded relationships with any specific outcome. Fourth, the algorithm simultaneously applies multiple risk windows to identify effects with differential time-to-event relationships, such as acute, subacute, insidious or delayed onset. Fifth, COMPASS produces a composite score based on adjusted risk differences and ratios that enable prioritization across multiple potential safety concerns based on both magnitude of effect and public health significance. Finally, in contrast to traditional pharmacoepidemiology evaluation designs, which are typically implemented to estimate the effect of one drug-condition pair, the COMPASS model is designed to be scalable to allow estimation of multiple drug-outcome pairs concurrently, and is computationally feasible to screen thousands of potential adverse events within hours. This efficiency enables key principles of pharmacoepidemiology to be brought to bear during the initial exploratory phase of hypothesis generation.

## Analysis

Drug-outcome effect estimates generated from all methods were compared to a binary classification made to partition the test cases into ‘positive controls’ and ‘negative controls’. The classification was performed by OMOP through systematic review of structured product labels available on the FDA website before December 19, 2009, using the occurrence of a condition in the adverse event section of the majority of labels within a class as a surrogate for a ‘positive control’, and selecting conditions unrelated to any labeled events as ‘negative controls’ (16). For ACE inhibitors, 84 ‘positive controls’ and 2780 ‘negative controls’ were identified and used for experimentation. The ‘positive controls’ include labeled events known to be related to ACE inhibitor exposure, such as cough, hypotension, hyperkalemia, and renal impairment (17). ‘Negative controls’ include a wide range of conditions observed in the database that are unrelated to any known effect of exposure, such as uterine leiomyoma, osteomyelitis, ankle fracture, incisional hernia, malignant neoplasm of brain, and hammer toe. The full set of test cases is available for download at (16). Sensitivity was measured as the proportion of the 84 labeled events identified at statistically significant levels, based on alpha = 0.05 and 0.001. Specificity was measured as the fraction of the 2780 negative controls that failed to meet statistical significance. Positive predictive value was estimated as the proportion of the outcomes meeting statistical significance that were classified as positive controls.

For each method, a receiver operating characteristic (ROC) curve was produced. All drug-outcome pairs were rank-ordered by the effect size point estimate, and the sensitivity and specificity was estimated at all observed threshold values. Five complementary measures of performance were estimated based on these ROC curves. The c statistic, or the area under the ROC curve, provides a predictive probability that two random drug-outcome pairs, one positive control and one negative control, would be properly rank-ordered. The c statistic ranges from 0 to 1, with 1 indicating perfect prediction and 0.5 a random prediction. Partial area under ROC curve at 10% false positive (PAUC10) is used to focus on the highest scores and eliminate the range of the ROC curve with unacceptable low specificity. The value of PAUC10 ranges from 0 to 0.10, with random prediction scoring 0.005. Recall at 5% false positive (RECALL5) estimates what fraction of the positive controls is identified at a threshold of 95% specificity. Precision at 100 (P100) provides a measure of what proportion of the drug-outcome pairs amongst the 100 highest estimates are positive controls. 'Mean average precision' (MAP) is a metric commonly used in information retrieval that provides the average precision at each threshold value that represents a 'true positive' association.

Each method has multiple parameter settings, based on design decisions around surveillance windows to define time-at-risk, covariates to include, and metrics to calculate. Parameter settings for all methods were selected by choosing the configuration that maximizes PAUC10. Sensitivity analysis was performed to assess impact of different design decisions of the all performance measures.

**Results**

Table 1 shows that COMPASS generates the fewest significant signals (n=114) of all methods, and also has the highest precision (0.31). COMPASS dominates DP, with both higher sensitivity and specificity. USCCS has a higher sensitivity than COMPASS (0.61 vs. 0.42), but comes at the expense of four-fold increase in false positive rate (0.13 vs. 0.03). OS has the highest sensitivity (0.71) but also the lowest specificity (0.55) and lowest precision.

Figure 2 shows the impact of changing the alpha threshold for statistical significance from a=0.05 to a=0.001 on the number of signals, sensitivity and specificity of all four methods. COMPASS continues to produce the fewest signals of all methods, but identifies 34 fewer significant associations at the stricter threshold. Decreasing the significance level from a=0.05 to a=0.001 decreases COMPASS sensitivity from 0.42 to 0.37, while increasing specificity from 0.97 to 0.98. COMPASS sensitivity and specificity remain higher than DP. For USCCS, specificity increases from 0.87 to 0.94, while sensitivity decreases to 0.50. OS identifies the same 60 true positives at both alpha levels, but the false positive rate decreases from 0.45 to 0.35. Under both alpha levels, COMPASS has the highest precision of the four methods, increasing to 0.39 for a=0.001. In other words, 39% of the 80 signals identified by COMPASS with p<0.001 were true labeled events.

Across all five summary measures, COMPASS has the best performance in classifying ACE inhibitor labeled events from negative controls (Table 2). Figure 3 highlights the receiver operating characteristics (ROC) curves for each method, used to derive these statistics. ROC curves do not reach 100% sensitivity and 100% false positive rate, because methods may fail to produce an estimate if no events are observed for a given outcome, in which case AUC calculations assume unscored conditions receive the minimum possible score. COMPASS observed the highest predictive value (c = 0.648), with the ROC curve furthest departed from random prediction shown with the dashed diagonal line. If unscored conditions were ranked at random instead of given minimum score, then c statistic would increase to 0.738. DP has the next-highest AUC, c=0.631. USCCS has the least predictive model, c=0.555. Each ROC is annotated with point estimate thresholds to facilitate comparison of interpreting observed scores from each method. When defining the threshold as relative risk (RR) > 1.0, COMPASS has sensitivity = 0.51 and specificity = 0.85. At the same threshold of requiring a positive effect, DP is observed to have sensitivity/specificity tradeoff of (0.48/0.76), OS is (0.75/0.36), and USCCS is (0.69/0.42). Imposing a stricter criteria that RR>1.4, COMPASS has (0.06/0.99), DP has (0.26/0.92), OS has (0.25/0.81), USCCS has (0.13/0.87). COMPASS does not identify any label events at RR>2, but DP has (0.13/0.97), OS is (0.25/0.81), and USCCS is (0.06/0.99). A threshold of RR>3.0 is required to observe a false positive rate < 0.05 for OS. The estimate distributions suggest each method has different degrees of positive bias, with OS and USCCS being most susceptible to generating positive effect estimates.

For COMPASS, the optimal setting set the washout period to 180d, specified inclusion criteria that at least one indication diagnosis is observed prior to index exposure, excluded all patients with contraindication in 30d prior first exposure, applied 20 propensity score strata, and identified incident events within 30d from exposure start. This setting produced the maximum PAUC10, MAP, and RECALL5. The maximum P100 was 0.15, observed when changing the washout period to 90d. The highest observed AUC for COMPASS was 0.673, by not restricting by prior indications, reducing the number of strata to 10, and including all events during the 30d post-exposure start.

For DP, the optimal setting selected prevalent events, used the empiric Bayes geometric mean metric from the MGPS algorithm, did not stratify on age and gender, and applied a surveillance window of 30d from exposure start. This setting also yielded the highest MAP, P100, and RECALL5 amongst the DP configurations. Changing the condition type parameter to incidence events produced a higher AUC (c=0.637). For OS, the setting that yielded the highest PAUC10 used first exposures to drugs, first occurrence of conditions, and compared the rate of events with all time post-exposure (excluding the index date) to the overall background rate. The maximum AUC amongst the OS configurations was 0.615, maximum MAP = 0.053, and maximum RECALL5=0.15, all obtained by using all drug exposures and all condition occurrences and restricting the surveillance window to 30d from exposure start, including the index date. The maximum P100 observed for OS was 0.11. For USCCS, the optimal setting selected incident events, excluded the index date of exposure, defined time-at-risk as the length of exposure + 60d, and specified the precision of the Normal prior as 2. This setting was the maximum self-controlled configuration based on P100 and RECALL5 as well. A different setting, using 30d from exposure start as the surveillance window, had the maximum AUC (c=0.680) and MAP = 0.054 for USCCS.

One potential explanation for performance differences between the methods is the extent to which confounding factors could influence results. Table 3 shows the impact that propensity score adjustment played within the COMPASS method. Relative risks for each adverse event were estimated by comparing the observed outcome rate in the exposed population (those with incident exposure to ACE inhibitors) to an unexposed population constructed as those patients with incident exposure to an alternative drug that shares the same indication as ACE inhibitors but has a different mechanism of action. The drug set identified included angiotensin receptor blockers, beta blockers, calcium channel blockers, and diuretics. Both cohorts were restricted to include patients with a recorded diagnosis for one of the ACE inhibitor indications, and excluded patients with a recent diagnosis of any ACE inhibitor contraindication. These two cohorts were observed to have important differences, with ACE inhibitor users having a higher proportion of males, greater medication use and procedures recorded, higher Charlson comorbidity index, with higher background rates of diabetes, congestive heart failure, hypertension, renal crisis scleroderma, and diabetic nephropathy. Covariate adjustment through propensity score stratification reduced the observed imbalanced to <5% differences between cohorts. Inherent in the USCCS approach is the self-controlled design that is intended to address time-invariant confounders, but temporal changes in health service utilization and increasing disease severity can bias results. Both DP and OS are unadjusted association measures, using observed rates from the overall population to calculate expectations to use to compare with observed counts, so these methods could be susceptible to bias from any of the covariate imbalances identified. DP and OS could also be biased by additional factors not observed in COMPASS due to the required similarities between the exposed and unexposed populations in the design.

Figure 4 highlights the observed effect estimates and 95% confidence intervals of all four methods for 35 selected labeled events. Among these labeled conditions, COMPASS identified 17 statistically significant associations, 11 of which were among the conditions identified with false positive rate < 0.05 (Cough, diarrhea, functional diarrhea, gout, heartburn, impaired renal function disorder, low blood pressure, nausea, orthostatic hypotension, pruritus of skin, and vomiting). For the same set of labeled events, DP also identified 17 statistically significant associations with 11 conditions having 5% false positive rate or less. Only three of the conditions identified were the same as those identified by COMPASS (gout, low blood pressure, and orthostatic hypotension). For OS, 29 of the labeled events were identified at $p<0.05$, but only eight of those conditions were detected with specificity > 95%. USCCS had 9 conditions reach statistical significance at a false positive rate lower than 5%.

Amongst the four methods, COMPASS was the only method to identify diarrhea, functional diarrhea, heartburn, nausea, pruritus of skin, and vomiting. USCCS was only method to identify bronchospasm, edema of larynx, and leukopenia. DP exclusively identified chest pain and palpitations, while OS was the only method to detect asthenia. Three events (acute laryngopharyngitis, asthma, pemphigus) were not identified as statistical significant by any method. Additionally, 10 conditions were not identified by any method with specificity > 95%: acquired hemolytic anemia, allergic urticaria, anemia, constipation, dyspnea, flushing, generalized abdominal pain, insomnia, oliguria and anuria, and primary thrombocytopenia.

## Discussion

We compared four different active surveillance methods in a retrospective claims database analysis to determine their ability to identify true safety findings and discern from negative control events within ACE inhibitor exposures. COMPASS was the best performing method in terms of fewest signals, lowest false positive rate, highest predictive probability, and greatest precision. If greater sensitivity is desired, the self-controlled case series design outperforms COMPASS but comes at expense of a four-fold increase in false positives. Disproportionality analysis is outperformed by COMPASS in terms of both sensitivity and specificity. Observational screening is positively biased with estimates consistently greater than relative risk > 1, and produces an unacceptable false positive rate at conventional levels of statistical significance.

COMPASS attempts to incorporate accepted practice in pharmacoepidemiology evaluation studies, with bioinformatics innovations to make automated heuristics for many of the design decisions typically customized on a case-by-case basis. The method is similar in intent to the high-dimensional propensity score (HDPS) approach proposed by Schneeweiss et al. (18), but differs in its covariate selection procedures. The HDPS heuristic requires both the exposure and outcome to identify empiric confounders, whereas COMPASS uses only information about the exposed and unexposed populations. Rubin argues that outcome information should not be used to estimate probability of treatment (19), but Brookhart et al. showed in simulation study the potential for a covariate related to exposure and unrelated to outcome to inflate variance without decreasing bias (20). From a technical standpoint, COMPASS is more scalable than HDPS for examining large sets of drugs and outcomes on a continual basis, because it only requires constructing one propensity score model for each exposure that can be applied across all outcomes. All COMPASS analyses of ACE inhibitors, 480 configurations executed for 2864 outcomes against the CCAE database, were run in the OMOP research lab in 91 hours. An additional advantage of COMPASS as an active surveillance tool is that its systematically applies existing clinical information to objectively and transparently define study design decisions (comparator selection, inclusion/exclusion criteria, covariate adjustment) and facilitates a comprehensive sensitivity analysis, rather than requiring expert assessment to customize a study plan that can be timely to develop and subject to potential disagreement across stakeholder groups.

Both the AUC and PAUC10 suggest that all four methods are better than random predictions, but each leave substantial room for improvement. While the COMPASS method approaches AUC in line with other clinical diagnostics in the range of 0.65-0.80 (21-26), in the context of an active surveillance system, the frequency of false positives and the potential for false negatives may be considered unacceptable. Here, it is important to reinforce the need for an active surveillance system to complement, rather than replace, existing practice and to be used to generate and prioritize potential hypotheses that require further evaluation. Also, while COMPASS is observed to have better overall performance than DP, OS, and USCCS, the event-specific analysis suggests that multiple methods may be necessary for an active surveillance to comprehensively evaluate the full spectrum of events, as each method uniquely identified positive controls. Further research is needed to determine why these patterns exist and for developing strategies to integrate evidence from across different methods.

We believe our results provide a useful first step toward characterizing the expected performance of an active surveillance system in its ability to reliably identify true drug safety issues. The chief limitation in our study is our focus on one drug class, as several factors could influence performance across medical products, including prevalence and duration of exposure, maturity of the drug class and clinical comfort with the mechanism of action, disease complexities in the underlying indicated population, and the potential for differential confounding across different safety effects. Further retrospective studies of other products would aid in increasing both the precision of the performance estimates and also improve the generalizability of the findings to support our use of the active surveillance system prospectively.

Another limitation of the current study is the potential for misclassification in the definition of ‘positive controls’ and ‘negative controls’. This ground truth classification was based on events occurring on the product label, but some adverse events may be listed on labels due to observations from clinical trials or spontaneous reports but in absence of definitive evidence of a true causal relationship. Similarly, negative controls were selected based on the condition being unrelated to any labeled event, though it is possible that there is a previously unknown association that has been uncovered that is instead being classified as a ‘false positive’ within this study. The risk of a true negative control seems minimal, since ACE inhibitors are mature products with presumably well-understood safety profiles. Further misclassification can arise due to the mapping of the labeled events to specific diagnosis codes that occur in the data. In this study, all outcomes defined by occurrence of diagnosis codes may limit method performance, and performance may be differential to the severity of the disease. In a typical pharmacoepidemiology evaluation study of a specific drug-outcome association, the outcome definition may be more refined to include sets of diagnosis codes, potentially in conjunction with diagnostic or treatment procedures and laboratory values. For an active surveillance system to be applied across a large array of outcomes, this health outcome of interest definitions would need to be developed a priori.

The current study used all data accumulated over time in CCAE, from 2003 to 2008. Since an active surveillance system may also be applied to newly marketed products with accumulating exposures in a continuous fashion, further analysis should evaluate method performance in that context. ACE inhibitors do not make a good case study for this scenario, since the class reflects a mature product that was approved prior to data availability in most data systems.  Thomson MarketScan Commercial Claims and Encounters database reflects only one viable data source that could contribute to an active surveillance network.  It reflects privately insured population, so may not be generalizable to the overall US population and may not accurately reflect ACE inhibitor use of adverse event experiences.  Administrative claims data reflects data captured for reimbursement, and may be different than data capture systems for electronic health records.  Method performance should be evaluated across multiple disparate sources.  Strategies for integrating estimates across the data network should also be explored.

**Tables and Figures**

Table 10: Operating characteristics of the four methods at alpha=0.05

| Method | Total signals | True positives | False positives | Sensitivity | Specificity | Precision |
|---|---|---|---|---|---|---|
| COMPASS | 114 | 35 | 79 | 0.42 | 0.97 | 0.31 |
| DP | 252 | 31 | 221 | 0.37 | 0.92 | 0.12 |
| USCCS | 402 | 51 | 351 | 0.61 | 0.87 | 0.13 |
| OS | 1302 | 60 | 1242 | 0.71 | 0.55 | 0.05 |

COMPASS-Comparator-adjusted Safety Surveillance; DP-Disproportionality analysis; USCCS-Univariate self-controlled case series; OS-Observational screening

Table 11: Performance measures of the four methods. AUC-Area under receiver operating characteristic curve; PAUC10 - Partial AUC at 10% false positive rate; MAP- mean average precision; P100- precision at top 100 signals; RECALL5 - recall at 5% false positives

| METHOD | AUC | PAUC10 | MAP | P100 | RECALL5 |
|---|---|---|---|---|---|
| COMPASS | 0.648 | 0.023 | 0.085 | 0.14 | 0.274 |
| DP | 0.631 | 0.017 | 0.075 | 0.12 | 0.202 |
| OS | 0.573 | 0.014 | 0.05 | 0.10 | 0.131 |
| USCCS | 0.555 | 0.011 | 0.044 | 0.09 | 0.131 |

Metrics: AUC-Area under receiver operating characteristic curve (min=0.0; random=0.5; max=1.0); PAUC10 - Partial AUC at 10% false positive rate (min=0.0; random=0.005; max=0.1); MAP- mean average precision (min=0; max=1); P100- precision at top 100 signals (min=0; max=0.84); RECALL5 - recall at 5% false positives (min=0; max=1)
Methods: COMPASS-Comparator-adjusted Safety Surveillance; DP-Disproportionality analysis; USCCS-Univariate self-controlled case series; OS-Observational screening

Table 12: COMPASS propensity score balance effects; Exposed- ACE inihibitor; Unexp- unexposed; RR- relative risk

| | **Pre-adjustment** | | | **Post-adjustment** | | |
|---|---|---|---|---|---|---|
| | **Exposed** | **Unexp** | **RR** | **Exposed** | **Unexp** | **RR** |
| **Demographics** | | | | | | |
| Age | 51.71 | 50.34 | 1.03 | 50.79 | 51.21 | 0.99 |
| Gender (% male) | 0.56 | 0.47 | 1.20 | 0.50 | 0.49 | 1.02 |
| **Lifestyle risk factors** | | | | | | |
| Smoking | 0.07 | 0.07 | 0.98 | 0.06 | 0.06 | 0.98 |
| Obesity | 0.06 | 0.05 | 1.27 | 0.05 | 0.06 | 0.93 |
| Alcohol | 0.05 | 0.04 | 1.13 | 0.06 | 0.04 | 1.34 |
| Drug abuse | 0.01 | 0.01 | 1.02 | 0.01 | 0.01 | 1.01 |
| **Health service utilization measures** | | | | | | |
| Total medication count | 1.30 | 1.20 | 1.09 | 1.24 | 1.22 | 1.02 |
| Indication medication count | 1.03 | 0.16 | 6.34 | 0.25 | 0.40 | 0.62 |
| Total procedure count | 5.92 | 5.64 | 1.05 | 5.94 | 5.59 | 1.06 |
| Total outpatient visits | 2.57 | 2.65 | 0.97 | 2.53 | 2.61 | 0.97 |
| Total inpatient visits | 0.41 | 0.40 | 1.03 | 0.37 | 0.38 | 0.97 |
| % Exposed within inpatient visit | 0.10 | 0.11 | 0.90 | 0.10 | 0.10 | 1.00 |
| **Charlson score** | 1.86 | 1.36 | 1.36 | 1.58 | 1.51 | 1.05 |
| **Comorbidities within Charlson index (% of persons with condition)** | | | | | | |
| Diabetes (mild to moderate) | 0.44 | 0.26 | 1.67 | 0.34 | 0.32 | 1.05 |
| Diabetes with chronic complications | 0.26 | 0.15 | 1.75 | 0.18 | 0.19 | 1.00 |
| Chronic pulmonary disease | 0.20 | 0.19 | 1.04 | 0.21 | 0.19 | 1.10 |
| Congestive heart failure | 0.11 | 0.07 | 1.70 | 0.06 | 0.08 | 0.84 |
| Cerebrovascular disease | 0.10 | 0.09 | 1.08 | 0.10 | 0.09 | 1.09 |
| Any malignancy | 0.06 | 0.07 | 0.92 | 0.06 | 0.07 | 0.91 |
| Peripheral vascular disease | 0.06 | 0.04 | 1.51 | 0.07 | 0.04 | 1.63 |
| Renal disease | 0.05 | 0.02 | 1.91 | 0.04 | 0.03 | 1.38 |
| Rheumatologic disease | 0.04 | 0.04 | 1.07 | 0.04 | 0.04 | 0.93 |
| Mild liver disease | 0.03 | 0.03 | 1.10 | 0.04 | 0.03 | 1.57 |
| Myocardial infarction | 0.02 | 0.01 | 2.73 | 0.01 | 0.01 | 0.76 |
| Peptic ulcer disease | 0.01 | 0.01 | 1.07 | 0.01 | 0.01 | 0.93 |
| Metastatic solid tumor | 0.01 | 0.02 | 0.61 | 0.01 | 0.01 | 0.95 |
| Dementia | 0.01 | 0.01 | 1.26 | 0.01 | 0.01 | 0.89 |
| Hemoplegia or paralegia | 0.01 | 0.01 | 0.90 | 0.01 | 0.01 | 0.97 |
| **Indication covariates (% of persons with condition)** | | | | | | |
| Hypertension | 0.76 | 0.70 | 1.08 | 0.72 | 0.76 | 0.95 |
| Hypertensive Emergencies | 0.52 | 0.43 | 1.22 | 0.47 | 0.49 | 0.97 |
| Renal Crisis Scleroderma | 0.51 | 0.42 | 1.21 | 0.46 | 0.48 | 0.97 |
| Hypertension due to Scleroderma | 0.50 | 0.41 | 1.22 | 0.45 | 0.46 | 0.97 |
| Diabetic Nephropathy | 0.29 | 0.17 | 1.65 | 0.21 | 0.21 | 1.00 |
| Myocardial Infarction | 0.24 | 0.27 | 0.89 | 0.24 | 0.24 | 0.97 |
| Myocardial Infarction Prevention | 0.22 | 0.25 | 0.88 | 0.22 | 0.23 | 0.97 |
| Prevention of Cerebrovascular Accident | 0.19 | 0.24 | 0.81 | 0.21 | 0.21 | 0.97 |
| Left Ventricular Dysfunction following Myocardial Infarction | 0.13 | 0.10 | 1.40 | 0.09 | 0.10 | 0.90 |
| Prevention of Recurrent Atrial Fibrillation | 0.11 | 0.09 | 1.20 | 0.10 | 0.09 | 1.09 |
| Chronic Heart Failure | 0.10 | 0.06 | 1.63 | 0.06 | 0.07 | 0.84 |
| Diastolic Heart Failure | 0.08 | 0.05 | 1.67 | 0.04 | 0.05 | 0.84 |
| Edema | 0.08 | 0.07 | 1.13 | 0.06 | 0.07 | 0.90 |
| Diabetic Retinopathy | 0.07 | 0.04 | 1.88 | 0.04 | 0.05 | 0.81 |
| Nondiabetic Proteinuric Nephropathy | 0.05 | 0.03 | 1.51 | 0.03 | 0.04 | 0.88 |
| Migraine Prevention | 0.05 | 0.06 | 0.82 | 0.05 | 0.05 | 0.99 |
| Cystine Renal Calculi | 0.04 | 0.04 | 1.13 | 0.04 | 0.04 | 0.95 |
| Raynaud's Phenomenon | 0.03 | 0.03 | 1.33 | 0.06 | 0.03 | 2.12 |
| Asymptomatic Left Ventricular Dysfunction | 0.02 | 0.02 | 1.25 | 0.02 | 0.02 | 0.90 |

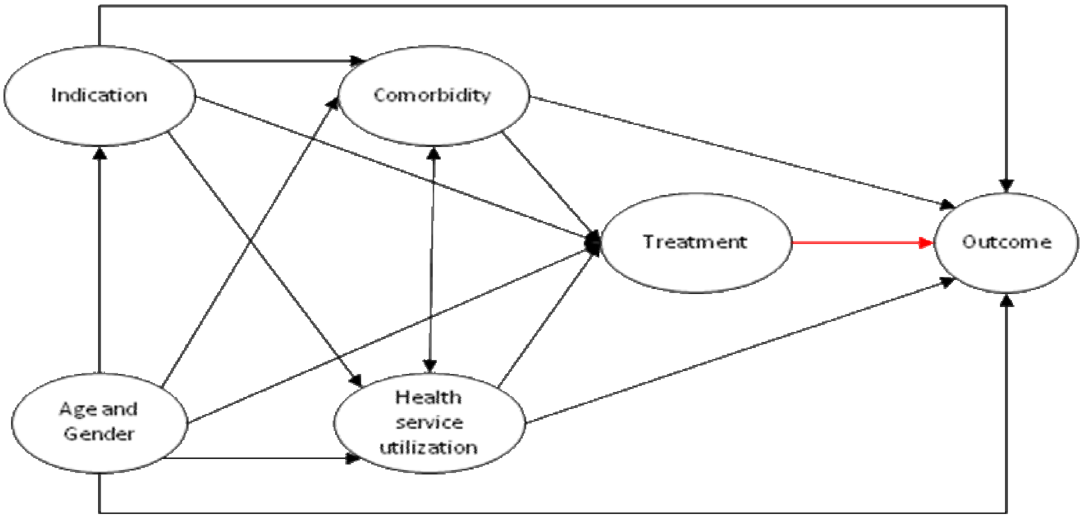


Figure 18: COMPASS conceptual model

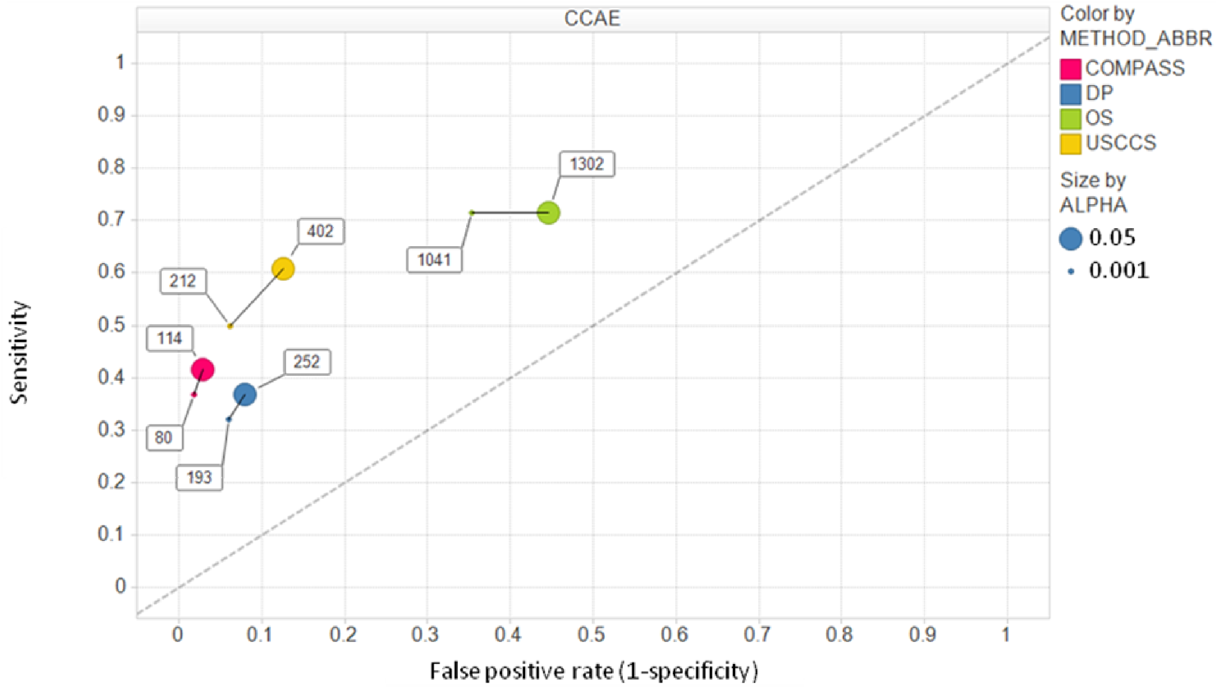


Figure 19: Sensitivity and specificity of four methods at alpha=0.05 and 0.001. Each method is annotated with the number of signals generated at that significance threshold.
Methods: COMPASS-Comparator-adjusted Safety Surveillance; DP-Disproportionality analysis; USCCS-Univariate self-controlled case series; OS-Observational screening

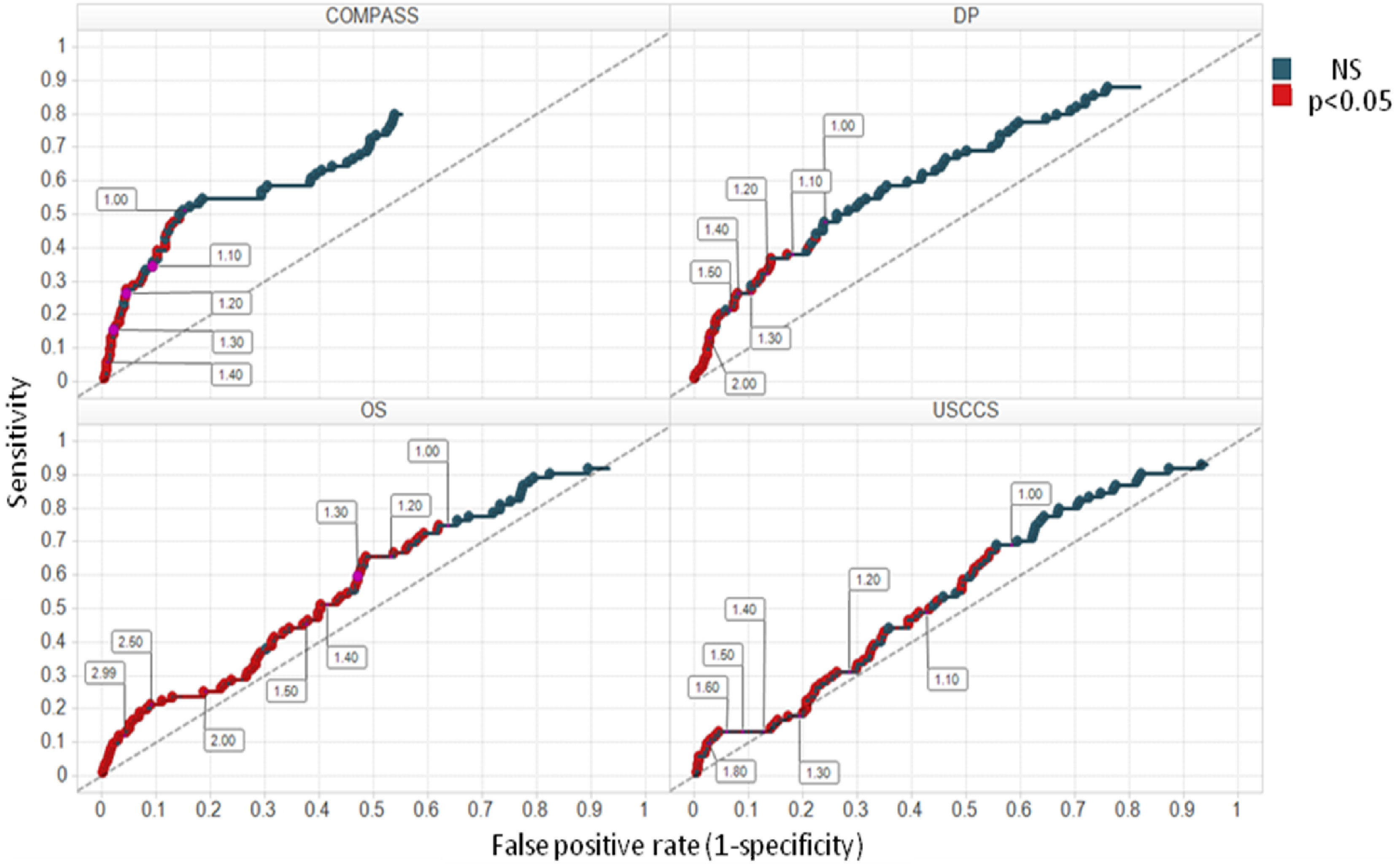


Figure 20: Receiver operating characteristics curves for each of the four methods. Select point estimate thresholds are annotated on each ROC curve to highlight sensitivity/specificity tradeoff at different observed effect sizes (relative risk).

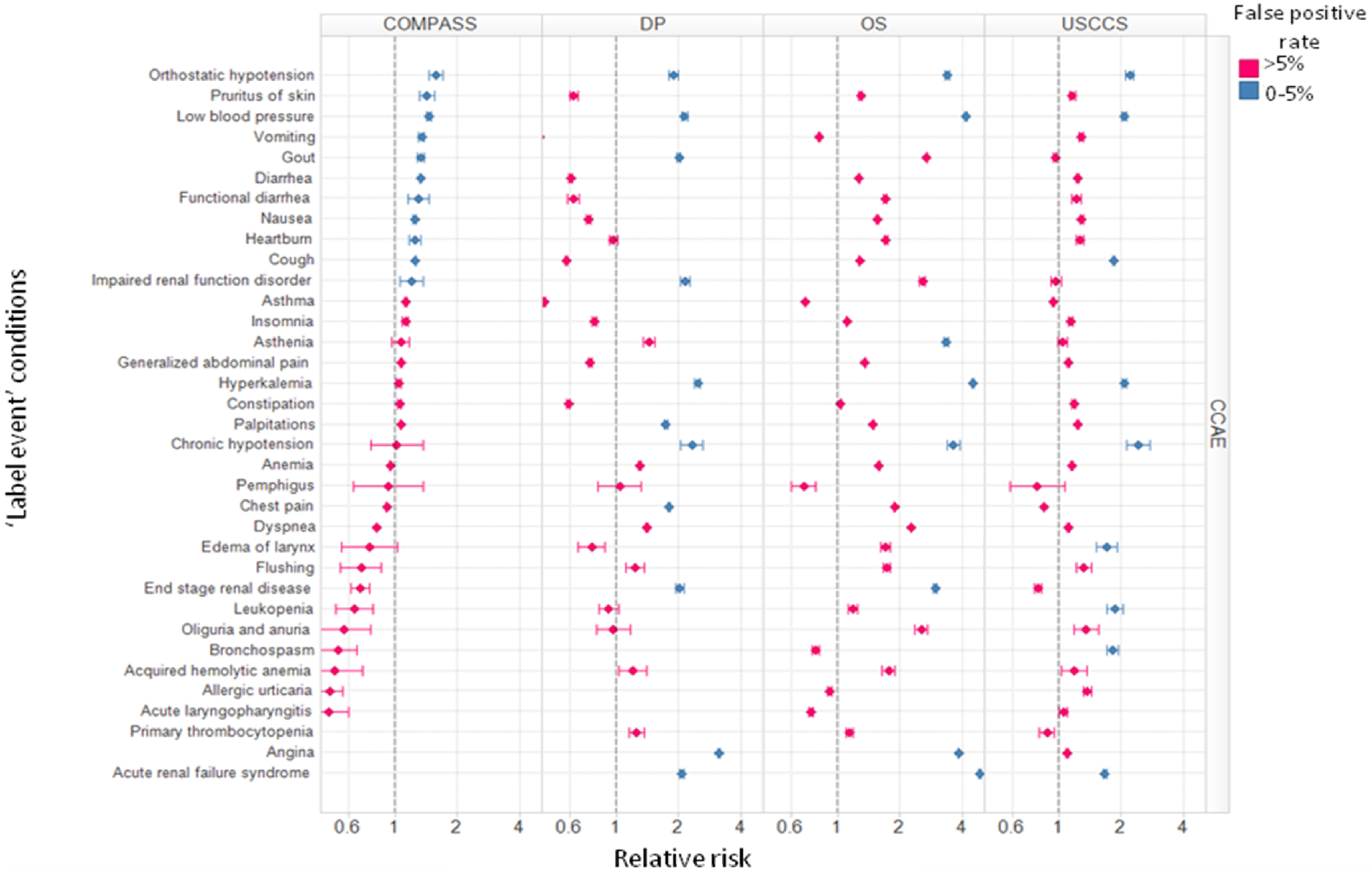


Figure 21: Estimates for 'label event' conditions across the four methods, ordered by estimate from COMPASS. Estimates are colored based on the false positive rate at the point estimate threshold.

## References

1. Berlin JA, Glasser SC, Ellenberg SS. Adverse event detection in drug development: recommendations and obligations beyond phase 3. *Am J Public Health.* Aug 2008;98(8):1366-1371.

2. Almenoff J, Tonning JM, Gould AL, et al. Perspectives on the use of data mining in pharmacovigilance. *Drug Saf.* 2005;28(11):981-1007.

3. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol.* Apr 2005;58(4):323-337.

4. Woodcock J, Behrman RE, Dal Pan GJ. Role of Postmarketing Surveillance in Contemporary Medicine. *Annu Rev Med.* Jan 2010.

5. FDA. The Sentinel Initiative: A National Strategy for Monitoring Medical Product Safety. May 2008; http://www.fda.gov/Safety/FDAsSentinelInitiative/ucm089474.htm. (Accessed January 3, 2011).

6. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med*. 2010 Nov 2;153(9):600-6.

7. Ryan PB. Review of Observational Analysis Methods. Observational Medical Outcomes Partnership. Foundation for National Institutes of Health. http://omop.fnih.org/?q=node/61. (Accessed January 3, 2011).

8. Nelson J, Cook A, Yu O. Evaluation of signal detection methods for use in prospective postlicensure medical product safety surveillance. http://www.fda.gov/OHRMS/DOCKETS/98fr/FDA-2009-N-0192-rpt.pdf. (Accessed January 3, 2011).

9. Ryan PB, Powell, G.E., Pattishall, E.N., Beach, K.J. Performance of Screening Multiple Observational Databases for Active Drug Safety Surveillance. *International Society of Pharmacoepidemiology*. Providence, RI; 2009.

10. Whitaker HJ, Hocine MN, Farrington CP. The methodology of self-controlled case series studies. *Stat Methods Med Res.* Feb 2009;18(1):7-26.

11. Whitaker HJ, Farrington CP, Spiessens B, Musonda P. Tutorial in biostatistics: the self-controlled case series method. *Stat Med.* May 30 2006;25(10):1768-1797.

12. Observational Medical Outcomes Partnership. Foundation for National Institutes of Health. OMOP Methods Library; 2011. (http://omop.fnih.org/MethodsLibrary). (Accessed January 3, 2011).

13. Beach KJ, Le HV, Powell G, Pattishall E, Ryan PB, Mera R. Performance of a Semi-Automated Method for Risk Estimation using Observational Databases; 2009.

14. Merrill GH, Ryan, P.B., Painter, J.L. Using SNOMED to Normalize and Aggregate Drug References in the SafetyWorks Observational Pharmacovigilance Project. *KR-MED*. Phoenix, AZ, USA; 2008.

15. Whitaker H. The self controlled case series method. *BMJ.* 2008;337:a1069.

16. Ryan PB. Defining a Reference Set for Evaluating the Performance of Active Surveillance Method. Observational Medical Outcomes Partnership. Foundation for National Institutes of Health. http://omop.fnih.org/OMOPWhitePapers. (Accessed January 3, 2011).

17. Chou R, Helfand M, Carson S. *Drug Class Review on Angiotensin Converting Enzyme Inhibitors* 1995.

18. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology.* Jul 2009;20(4):512-522.

19. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*. Oct 1997;127(8 Pt 2):757-63.

20. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *Am J Epidemiol.* Jun 15 2006;163(12):1149-1156.

21. Folsom AR, Chambless LE, Ballantyne CM, Coresh J, Heiss G, et al. An assessment of incremental coronary risk prediction using C-reactive protein and other novel risk

markers: the atherosclerosis risk in communities study..*Arch Intern Med.* 2006 Jul 10;166(13):1368-73.

22. Ebell MH, Smith MA, Barry HC, Ives K, Carey M. The rational clinical examination. Does this patient have strep throat? *JAMA*. 2000 Dec 13;284(22):2912-8.

23. Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med*. 2005 Oct 27;353(17):1773-83.

24. Martin BJ, Finlay JA, Sterling K, Ward M, Lifsey D. Early detection of prostate cancer in African-American men through use of multiple biomarkers: human kallikrein 2 (hK2), prostate-specific antigen (PSA), and free PSA (fPSA). *Prostate Cancer Prostatic Dis*. 2004;7(2):132-7.

25. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med.* 2002 May 15;21(9):1237-56.

26. Bineau S, Dufouil C, Helmer C, Ritchie K, Empana JP, et al. Framingham stroke risk function in a large population-based cohort of elderly people: the 3C study. *Stroke.* 2009 May;40(5):1564-70.

# CHAPTER FIVE: MANUSCRIPT 2:
# “Integrating active drug safety surveillance analyses across a network of observational healthcare databases”

## Abstract

**Background:** The development of an active drug safety surveillance system requires access to a network of disparate observational healthcare data sources. There is little empirical evidence to anticipate performance of active surveillance analyses in their ability to identify true drug safety issues and discern from false positive findings or the consistency of evidence observed across data sources.

**Objectives:** To measure the operating characteristics of an active surveillance method in five disparate observational databases by retrospective evaluation of known adverse events associated with ACE inhibitor exposure.

**Results:** In all five databases, Comparator-Adjusted Safety Surveillance provided a moderately predictive model with high specificity > 97%. The total number of events reaching statistical significance and the sensitivity in identifying labeled events varied considerably by data source. Composite summaries based on meta-analysis of source-specific effect estimates did not yield additional predictive ability or identify additional outcomes not found using individual sources alone. 82% of the outcomes with a statistically significant composite effect estimate were observed to have high heterogeneity ($I^2$ statistic > 75%) of point estimates among databases.

**Conclusions:** Active surveillance across a network of disparate data sources can provide valid information to complement existing evidence as part of a comprehensive drug safety assessment. Independent replication of statistically significant findings improves precision of observational analyses, but does not eliminate risk of false positive findings. Substantial heterogeneity across data sources requires the development of a strategy to assess emerging drug safety issues by examining both source-specific effect estimates and composite summaries.

**Key words:** Active surveillance; Meta-analysis; Drug safety

## Introduction

In 2007, Congress passed the Food and Drug Administration (FDA) Amendment Act, which called for the establishment of an "active postmarket risk identification and analysis system" with access to patient-level observational data from 100 million lives by 2012[1]. In the US, creating such a system requires establishing a network of disparate data sources, as it is recognized that currently no single data holder has adequate capture of information from throughout the healthcare delivery system or sufficient sample that is adequately representative of the general population.

Several initiatives demonstrate the feasibility of coordinating a network of disparate observational data sources. The HMO Research Network has established a consortium of 16 health maintenance organizations to conduct multicenter pharmacoepidemiologic evaluation studies across their administrative claims data[2]. FDA's Mini-Sentinel pilot project announced access to data for 60 million persons, with current focus on administrative claims from privately insured populations[3]. As part of its methodological research efforts, the Observational Medical Outcomes Partnership (OMOP) established a data network of 10 data sources covering over 200 million persons, and representing the breadth of disparate data available (administrative claims aggregated from insurers, large employers, and directly from the point-of-care and electronic health records from inpatient systems, outpatient services, and across an integrated health information exchange) and the diversity of population demographics of interest[4].

In the context of active drug safety surveillance, where the data network is envisioned to be used for systematic monitoring of any medical product and any health outcome of interest, several methodological issues require careful consideration. Paramount to these pursuits is a full understanding of the accuracy of active surveillance methods that correctly identify true drug safety issues and discern false positive findings. Operating characteristics of methods, such as sensitivity, specificity and positive predictive value, may be influenced by attributes of the underlying data, including population size, patient demographics and health profile, completeness in data capture, and longitudinality of coverage, i.e. the ability to track all interactions with the healthcare system for individual patients over time within a dataset. Across a network of disparate data sources, method performance and the relative confidence in information gained from each contributing source may vary. There is a need to investigate how effect estimates from disparate sources can be meaningfully integrated to produce composite summaries, such as within a meta-analytic framework, and to assess the predictive performance of these summary estimates. It is also important to understand the extent of heterogeneity that may be present across different sources to help provide context and facilitate the proper interpretation of active surveillance results.

This study evaluated the performance of one active surveillance method, Comparator-Adjusted Safety Surveillance (COMPASS), across five disparate observational databases in its ability to identify known adverse events associated with ACE inhibitors exposure. The study also examined the operating characteristics of composite estimates and measured the heterogeneity across the data network.

## Methods

Five observational data sources from within the OMOP data network were used for this analysis. Thomson Reuters MarketScan Commercial Claims and Encounters (CCAE) is a large administrative claims database containing 59 million privately insured lives, and provides patient-level de-identified data from inpatient and outpatient visits and pharmacy claims of multiple large employers. MarketScan Lab Database (MSLR) contains 1.5 million persons representing a largely privately-insured population, with administrative claims from inpatient, outpatient, and pharmacy services supplemented by laboratory results. MarketScan Medicaid Multi-State Database (MDCD) provides administrative claims data for 11 million Medicaid enrollees from multiple states. MarketScan Medicare Supplemental and Coordination of Benefits Database (MDCR) captures administrative claims for 5 million retirees with Medicare supplemental insurance paid for by employers, including services provided under the Medicare-covered payment, employer-paid portion, and any out-of-pocket expenses. GE Centricity electronic health record (GE) contains patient-level data for 11 million persons captured at the point of care from a consortium of providers using the GE Centricity system in their outpatient and specialty practices. Analyses were conducted independently in each database, despite the potential for patient overlap across multiple databases, due to lack of unique patient identifier.

COMPASS is a statistical algorithm that estimates adjusted rate ratios for all outcomes of interest for a given medical product through propensity score stratification across exposed and unexposed cohorts for a given medical treatment. COMPASS applies a retrospective inception cohort design to compare the effects of the target drug to the effects of alternative treatments for the same indication defined by an automated heuristic. The COMPASS model focuses on minimizing bias from four primary sources: personal demographics (such as age and gender), confounding by indication, effects of comorbidity, and health service utilization. COMPASS follows many of the design features advocated in the literature[5, 6], and has been demonstrated to have better performance characteristics than other active surveillance methods under consideration, such as self-controlled case series, disproportionality analysis, and observational screening, when examining known effects within a large claims database[7]. COMPASS was executed against all data sources using the same configuration, specifically, having set the washout period (time from observation start to index exposure) to 90 days, specified inclusion criteria that at least one indication diagnosis was observed prior to index exposure, excluded all patients with contraindication in 30 days prior first exposure, applied 20 propensity score strata, and identified incident events within 30 days from exposure start. Complete details of the COMPASS method and implementation are available[7].

The Aniotensin Converting Enzyme (ACE) inhibitor drug class was selected for the retrospective evaluation of method performance, because of the products' widespread use and length of time on the market. Product longevity was important for accurately characterizing the product's existing safety profile, which may not be fully understood for newly marketed medicines[8]. Effect estimates generated from COMPASS were compared to a binary classification made to partition the test cases into 'positive controls' and 'negative controls'. The classification was performed by OMOP through systematic review of structured product labels available on the FDA website before December 19, 2009, using the occurrence of a condition in the adverse event section of the majority of labels within a class as a surrogate for a 'positive control', and selecting conditions unrelated to any labeled events as 'negative controls'[9]. For ACE inhibitors, 84 'positive controls' and 2780 'negative controls' were identified and used for experimentation. The 'positive controls' include labeled events known to be related to ACE inhibitor exposure, such as cough, hypotension, hyperkalemia, and renal impairment[10]. 'Negative controls' include a wide range of conditions observed in the database that are unrelated to any known effect of exposure, such as uterine leiomyoma, osteomyelitis, ankle fracture, incisional hernia, malignant neoplasm of brain, and hammer toe. The full set of test case outcomes is available for download from the OMOP website[9].

Effect estimates for all test cases were produced using COMPASS within each data source. Additionally, composite estimates summarizing the relative risks across the sources were produced using both fixed-effects and random-effects meta-analysis[11]. We used both meta-analytical approaches in order to assess the impact of heterogeneity on the accuracy of the composite estimates. Drug-outcome pairs were classified by how many sources produced significant associations to assess the impact of replicated findings on performance characteristics. Five operating characteristics were measured for each data source as well as the composite meta-analysis estimates. Sensitivity was measured as the proportion of the 84 labeled events identified at statistically significant levels, based on alpha = 0.05. Specificity was measured as the fraction of the 2780 negative controls that failed to meet statistical significance at alpha = 0.05. Positive predictive value (PPV) was estimated as the proportion of the outcomes meeting statistical significance that were classified as positive controls (e.g. labeled events). The c statistic, or the area under the Receiver Operating Characteristic (ROC) curve, provides a predictive probability that two random drug-outcome pairs, one true relation and one negative control, would be properly rank-ordered with the higher score being more likely to be true. The c statistic ranges from 0 to 1, with 1 indicating perfect prediction and 0.5 a random prediction. Partial area under the ROC curve at 10% false positive (PAUC10) is used to focus on the highest scores and eliminate the range of the ROC curve with unacceptably low specificity. The value of PAUC10 ranges from 0 to 0.10, with random prediction scoring 0.005. Heterogeneity across sources was measured using the $I^2$ statistic, classified as low ($I^2$<25%), medium, or high ($I^2$>75%) heterogeniety[12].

### Results

Table 13 provides a comparison of the source populations and data availability from across the data network used for this study. All databases include a higher proportion of females, with MSLR having the largest difference and CCAE having more balance. The databases had substantial variability in age distributions, with MDCD a greater number of younger persons and MDCR predominantly elderly. MDCD had the highest turnover rate, and GE had greater variability in the observation duration. Differences in the numbers of drugs, conditions, and procedures reflect the underlying disease severity of the source populations as well as the characteristics of the data capture process within each system. Within CCAE, there were over 3 million patients with at least one exposure to an ACE inhibitor. Restriction to incident use, defined as first exposure to an ACE inhibitor at least 180 days after observation period start, yielded over 1 million persons overall. The total sample size varied across the network of databases, but CCAE was the largest database and as such had more ACE inhibitor users. However, compared to the privately-insured population (reflected in CCAE), the proportions of ACE inhibitor users in the Medicare and

GE populations were markedly higher, likely as a result of the higher burden of illness in the elderly population.

The operating characteristics of COMPASS across the five data sources and amongst the composite meta-analysis estimates are presented in Table 14. CCAE produced the largest number of significant associations amongst the data sources (n=127), 38 of which were labeled events and 89 of which were negative controls, yielding both the highest sensitivity (0.45) and lowest specificity (0.97). CCAE had the highest AUC (0.645), and the second-largest PAUC10 behind GE. COMPASS identified only 3 significant associations within GE, all of which were true labeled events (cough, dysthymia, and shoulder joint pain) for a PPV of 1. All sources except GE had PPVs ranging from 0.30 to 0.36. MDCD and MDCR, despite representing disparate populations, shared 11 true positive findings in common and had very similar operating characteristics on all measures.

Ninety-four outcomes were statistically significant under the fixed-effects meta-analysis model, identifying two fewer true positives and 31 fewer false positives than CCAE alone. Estimates from the fixed effects model had strong correlations with CCAE estimates, since CCAE is the largest database and therefore has a higher weight, since weight is proportional to the inverse of the variance of the effect estimate within a database. As a result, operating characteristics between the fixed-effects and CCAE were similar, although the fixed-effects model had higher precision and PAUC10. The random-effects model had fewer outcomes reach significance than the fixed-effects model (n=14), with 57% of those events being true positives. The random-effects precision was higher than both the fixed-effect model and all individual sources, except for GE. The fixed-effects model had a comparable AUC with CCAE, while the random-effects meta-analysis yielded a less predictive model than individual AUC predictions from CCAE, MSLR, MDCR, and MDCD.

Across the five databases, 168 outcomes were identified as statistically significant in at least one source, with 48 of those outcomes being significant in 2 or more sources, 18 events in 3 or more sources, and 3 events in 4 sources (Table 14). No events were statistically significant in all five databases. Using significance in multiple sources as a criterion, we see that requiring 2 or more sources yields sensitivity of 0.37, specificity of 0.99, and PPV of 0.48. Requiring a majority (3 or more) of the sources to show a significant finding substantially increases precision to 0.72, as 13 of the 18 identified outcomes were true labeled events.

Two labeled events were identified in all but one database: cough (not MSLR) and diarrhea (not GE), but one negative control was consistently identified as a false positive in four databases: benign neoplasm of the colon (CCAE, MSLR, MDCD, MDCR). The combination of databases that produced consistent findings differed by outcome; amongst the 11 labeled events identified by 3 databases, 9 were found within CCAE, MDCD, and MDCR (including orthostatic hypertension, nausea, vomiting, insomnia, and arthralgia of the pelvic region), but dysthymia was identified in CCAE, MDCR, and GE, and shoulder joint pain was significant in CCAE, MDCD, and GE. Figure **22** provides forest plots for 42 labeled events, and highlights how the consistency in estimates across sources varies substantially by outcome.

Within the negative controls, all sources except GE identified false positive events that were not replicated in any other source. CCAE had 65 unique false positives, including hammer toe and multiple sclerosis. MSLR was the only source to falsely identify the two events, hemoglobin SS disease with crisis and mononeuritis of lower limb. MDCD had 24 false positives, including Acquired deformity of toe and candidiasis of the esophagus, while MDCR identified 15 such unique outcomes, including false positives for primary malignant neoplasm of vermillon border of lower lip and acquired spondyloisthesis. In no cases did meta-analysis (fixed or random-effects) produce a statistically significant estimate for a true positive or negative control that wasn't otherwise identified by at least one database individually.

Figure 23 highlights the magnitude of heterogeneity observed across all test cases, classified by both their status as a labeled event or negative control and also the statistical significance of the fixed-effects composite estimate. Amongst the 36 significant labeled events, 31 (86%) had $I^2$ values > 75%, indicating high heterogeneity. 79% of the 58 significant false positives were also observed to have $I^2$ values > 75%. In contrast, 24% of the 34 false negatives have high heterogeneity, and 15 of the outcomes have $I^2$ < 25%. 84% of the true negatives have low heterogeneity ($I^2$ <=25%), and only 4% of true negatives were observed with $I^2$ > 75%. There is no significant correlation between effect size and $I^2$ within each quadrant, though extremely small relative risks observed (RR<0.4) in the true negatives were associated with high heterogeneity.

**Discussion**

We evaluated an active surveillance method across a network of five observational databases to assess method performance in its ability to properly classify true drug safety issues from negative controls in a range of different data sources. In all five databases, COMPASS produced a moderately predictive model with high specificity > 97%. The total number of events reaching statistical significance and the sensitivity in identifying labeled events varied considerably by data source. With the exception of the GE data, all data sources had precision estimates suggesting that about one in three significant outcomes were labeled events.

The number of associations identified appears to be only marginally related with population size, as CCAE is 5 times larger than MDCD and 12 times larger than MDCR, but only yielded 2.5 times as many significant outcomes. This may be explained, in part, by the underlying source characteristics as the privately insured population in CCAE may be generally healthier, with fewer comorbidities and concomitant medications, and have less frequent health service utilization.

GE showed a notably different performance profile from the other databases; despite being the second-largest database in population size, it yielded the fewest significant outcomes (n=3) all of which were true positives. GE had the highest PAUC10 but the lowest AUC, in part due to the large number of outcomes for which COMPASS failed to generate estimated as a result of absence of observed cases of the drug-outcome co-occurrence. Unlike the other four sources, which derive drug exposure from pharmacy dispensing and condition occurrence from diagnosis codes on medical claims, GE drug exposure is inferred from prescriptions written and medication history and outcomes are identified from problem lists, both of which are generally maintained in outpatient centers and under-represent inpatient care. This finding suggests special attention is needed to understand the process that results in data capture, as analysis approaches based on assumptions of how claims data are captured may not generalize.

Composite estimates produced through conventional meta-analysis approaches (both fixed-effects and random-effects models) were not more predictive than estimates from individual sources, suggesting that pooling data from across sources will not necessarily provide greater confidence in assessing drug-outcome relationships. Meta-analysis has been a popular framework for aggregating effect estimates from multiple sources. In the context of randomized clinical trials, where effect estimates are assumed to be unbiased measures of the average treatment effect, conventional meta-analytical approaches need only be concerned with the variance within each effect estimate and the heterogeneity across estimates[13]. In contrast, observational analyses may be subject to biases, which may vary by data source due to the underlying data capture mechanism, such as accuracy of capture of measurable covariates and degree of unmeasured confounding. As a result, composite estimates produced from meta-analysis across a network of observational databases may present a false sense of precision, as meta-analytic methods do not address the nature or magnitude of bias that exists within each source and can't overcome the heterogeneity that exists across the data network. This study provides a first empirical evaluation of the magnitude of this potential problem and its impact on the predictive value of meta-analyses for exploring drug safety effects. These empirical results seem consistent with general guidelines for the use of meta-analysis that have been discussed previously in other contexts [14-17].

This study also demonstrates the relative importance of independent replication in observational analyses. Amongst outcomes identified as significant in only one source, 13% were labeled events. The precision increased to 33% when evaluating outcomes identified as significant in two sources, and increased substantially to 73% if outcomes were significant in three databases. These findings suggest that independent replication of a statistically significant association in multiple sources can be more informative than a single significant finding that is not substantiated in other sources, but observing significance in multiple sources does not eliminate the risk of false positive findings and does increase the risk of false negatives. Benign neoplasm of the colon, which was a negative control that showed significant associations in four databases, bears particular consideration. There has been an active debate about the potential merits of ACE inhibitors as treatment to prevent colorectal cancer, with some studies showing no association[18] and others hypothesizing a protective effect[19]. As a result, increased effects observed in this study could be the result of channeling bias and confounding by off-label indication, as patients who are at greater risk of having prior diagnosis for colon cancer may be more likely to be exposed to ACE inhibitors. This case study underscores the expected challenges to be faced by an active surveillance system and the need to evaluate the information produced from such a system in context with all other available evidence as part of a comprehensive safety assessment.

Another key finding from this study is the magnitude of the observed heterogeneity that is observed within effect estimates across the five data sources. 82% of the outcomes with a statistically significant composite effect estimate were observed to have high heterogeneity, with $I^2 > 75\%$. The results indicate that elevated risks identified within a network of databases are more likely to be accompanied by greater risk of variability in estimates across the network than drug-outcome pairs without observed relationships. The substantial heterogeneity explains, in part, the differences in performance observed between the fixed-effects and random-effects models, as fewer drug-outcome pairs reached significance within the random-effects model while the fixed-effects model was heavily weighted toward the largest database, CCAE. It is important to highlight the heterogeneity observed is based on variability between sources, but it can be difficult to determine the specific attribute within a source that is causing the inconsistent results. Further research, including the use of meta-regression techniques, may lend insight but may already require a larger network of data sources to compare. One desirable aspect of a network-based approach is that the central coordinating center can ensure estimates are received from across the network, and minimize the risk of bias due to differential reporting. Here, we've demonstrated the examination of forest plots to observe heterogeneity across sources, but funnel plots of sample size and variance can also provide a useful tool for examining bias across the reported estimates [20, 21].

Given the observed potential for large heterogeneity and the substantial variability in heterogeneity across outcomes, we recommend that active surveillance system results be presented with source-specific estimates in conjunction with any composite estimates, as typically shown in a forest plot. Pooling patient-level data across sources or statistical adjustment by adding source as a covariate (i.e., a set of indicator variables) in a multivariate model are unlikely to fully address the heterogeneity that can be present and may risk biased estimates leading to misinterpretation of the drug-outcome relationship. Guidelines for appropriate reporting of meta-analysis of observational studies provide a useful framework that could be followed[22].

A key limitation of this current study is the potential for misclassification in the definition of 'positive controls' and 'negative controls'. Ground truth was based on using the proxy of events occurring on the product label, but some adverse events may be listed on labels due to observations from clinical trials or spontaneous reports but in absence of definitive evidence of a true causal relationship.  Similarly, negative controls were selected based on the condition being unrelated to any labeled event, though it is possible that there is a previously unknown association that has been uncovered that is instead being classified as a 'false positive' within this study.  Further misclassification can arise due to the mapping of the labeled events to specific diagnosis codes that occur in the data, and the lack of confirmation of those event definitions through source record verification.  Method performance could be improved with greater precision in outcome definitions and reference set classification.

We believe our results provide a useful first step toward characterizing the expected performance of active surveillance analysis across a network of disparate observational databases in its ability to reliably identify true drug safety issues.  The chief limitation in our study is our focus on one drug class, as several factors could influence performance of both the method and the data sources across medical products, including prevalence and duration of exposure, maturity of the drug class and clinical comfort with the mechanism of action, disease complexities in the underlying indicated population, and the potential for differential confounding across different safety effects.  Another limitation of this study is the focus on one active surveillance method, as other approaches may exist or could be developed.  In that regard, these results could serve as a minimum benchmark to foster further methods innovation and evaluation.  Similarly, while the five data sources used in this analysis reflect the broad diversity of data available, additional data sources under consideration for inclusion in a national system should be evaluated.  The sources included in this study represent various populations of interest with different demographics and health behaviors (privately insured, Medicaid young, Medicare elderly) as well as both primary data capture processes (administrative claims and electronic health records), but the observed variability in performance suggest that operating characteristics are unlikely to be generalizable across databases and that each new data source needs to be assessed independently prior to inclusion in an active surveillance network.  Further retrospective studies of an array of drugs using a portfolio of alternative methods against a broader network of potential data sources would improve the applicability of the findings to support our use of the active surveillance system prospectively.

## Tables and Figures

Table 13: Data source characteristics

| | CCAE | MSLR | MDCD | MDCR | GE |
|---|---|---|---|---|---|
| Population (N) | N=59,836,290 | N=1,466,617 | N=11,188,360 | N=4,655,736 | N=11,216,208 |
| Gender | | | | | |
| Male: N (%) | 29,173,105 (48.75) | 515,174 (35.13) | 4,665,014 (41.70) | 2,071,968 (44.50) | 4,751,444 (42.36) |
| Female: N (%) | 30,663,185 (51.25) | 951,443 (64.87) | 6,523,346 (58.30) | 2,583,768 (55.50) | 6,460,828 (57.60) |
| Age (yrs) | | | | | |
| Mean (SD) | 32.4 (18.1) | 39.1 (17.5) | 23.4 (22.7) | 74.5 (8.0) | 40.6 (22.0) |
| Observation period length (mo) | | | | | |
| Mean (SD) | 21.2 (18.6) | 18.7 (11.1) | 14.2 (13.8) | 31.9 (22.9) | 24.0 (31.3) |
| Number of drug exposure records per person | | | | | |
| Median (25-75 %tile) | 9 (3-28) | 14 (5-35) | 14 (5-38) | 60 (20-134) | 8 (3-22) |
| Number of condition occurrence records per person | | | | | |
| Median (25-75 %tile) | 15 (5-39) | 27 (12-56) | 24 (9-63) | 57 (20-129) | 5 (2-10) |
| Number of procedure occurrence records per person | | | | | |
| Median (25-75 %tile) | 20 (7-52) | 39 (19-77) | 31 (12-70) | 72 (26-154) | 10 (3-24) |
| Any ACE Inhibitor exposure | | | | | |
| Prevalent users: N (%) | 3,052,264 (5.10) | 108,869 (7.42) | 614,703 (5.49) | 1,569,765 (33.72) | 1,361,058 (12.13) |
| Incident users: N (%) | 1,137,211 (1.90) | 32,532 (2.22) | 188,224 (1.68) | 483,853 (10.39) | 529,767 (4.72) |

CCAE: Thomson Reuters MarketScan Commercial Claims and Encounters
MSLR: Thomson Reuters MarketScan Lab Supplemental
MDCD: MarketScan Multi-state Medicaid database
MDCR: MarketScan Medicare Supplemental and Coordination of Benefits Database
GE: GE Centricity
Incident users, defined as first exposure >180 days from observation period start

Table 14: Operating characteristics of COMPASS across data sources and within composite summaries

| Source | Total signals | True positives | False Positives | Sensitivity | Specificity | PPV | AUC | PAUC10 |
|---|---|---|---|---|---|---|---|---|
| CCAE | 127 | 38 | 89 | 0.45 | 0.97 | 0.30 | 0.645 | 0.022 |
| MSLR | 9 | 3 | 6 | 0.04 | 1.00 | 0.33 | 0.598 | 0.014 |
| MDCR | 47 | 17 | 30 | 0.20 | 0.99 | 0.36 | 0.613 | 0.011 |
| MDCD | 51 | 16 | 35 | 0.19 | 0.99 | 0.31 | 0.608 | 0.012 |
| GE | 3 | 3 | 0 | 0.04 | 1.00 | 1.00 | 0.537 | 0.033 |
| | | | | | | | | |
| Meta-analysis composite estimates | | | | | | | | |
| Fixed effects | 94 | 36 | 58 | 0.43 | 0.98 | 0.38 | 0.644 | 0.032 |
| Random effects | 14 | 8 | 6 | 0.10 | 1.00 | 0.57 | 0.557 | 0.017 |
| | | | | | | | | |
| Threshold based on number of sources meeting significance | | | | | | | | |
| 1+ | 168 | 39 | 129 | 0.46 | 0.95 | 0.23 | | |
| 2+ | 48 | 23 | 25 | 0.27 | 0.99 | 0.48 | | |
| 3+ | 18 | 13 | 5 | 0.15 | 1.00 | 0.72 | | |
| 4+ | 3 | 2 | 1 | 0.02 | 1.00 | 0.67 | | |

CCAE: Thomson Reuters MarketScan Commercial Claims and Encounters
MSLR: Thomson Reuters MarketScan Lab Supplemental
MDCD: MarketScan Multi-state Medicaid database
MDCR: MarketScan Medicare Supplemental and Coordination of Benefits Database
GE: GE Centricity
‘Signal’- an outcome with a statistically significant association ($p<0.05$)
PPV: Positive predictive value
AUC: Area under receiver operating characteristic curve
PAUC10: Partial area under receiver operating characteristic curve, at 10% false positive rate

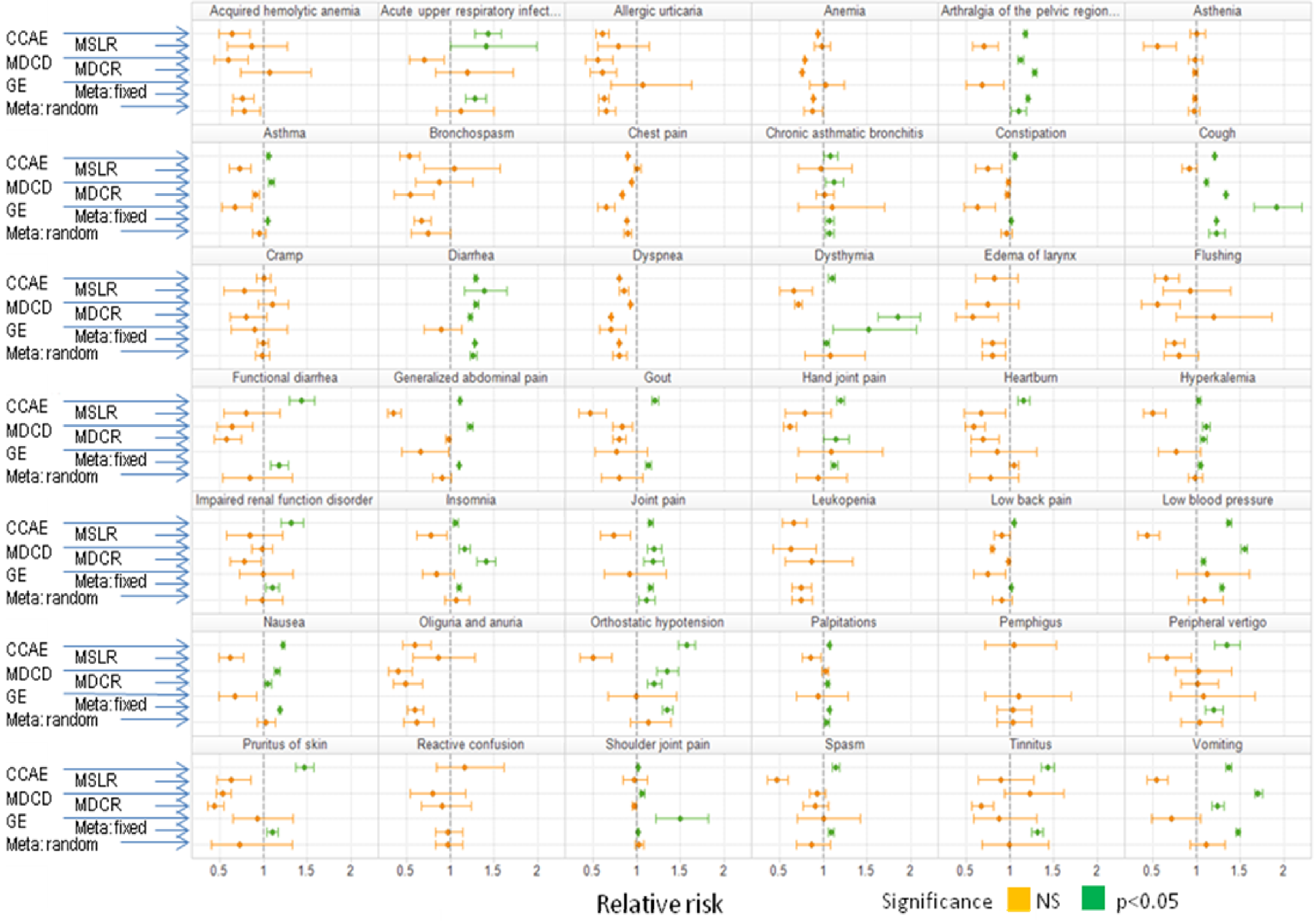


Figure **22**: Forest plots of effect estimates for 42 ACE inhibitor labeled events
Only positive relationships (RR>1) are highlighted for purposes of identification of risks.
Outcomes without estimates for specific databases are due to small case counts.

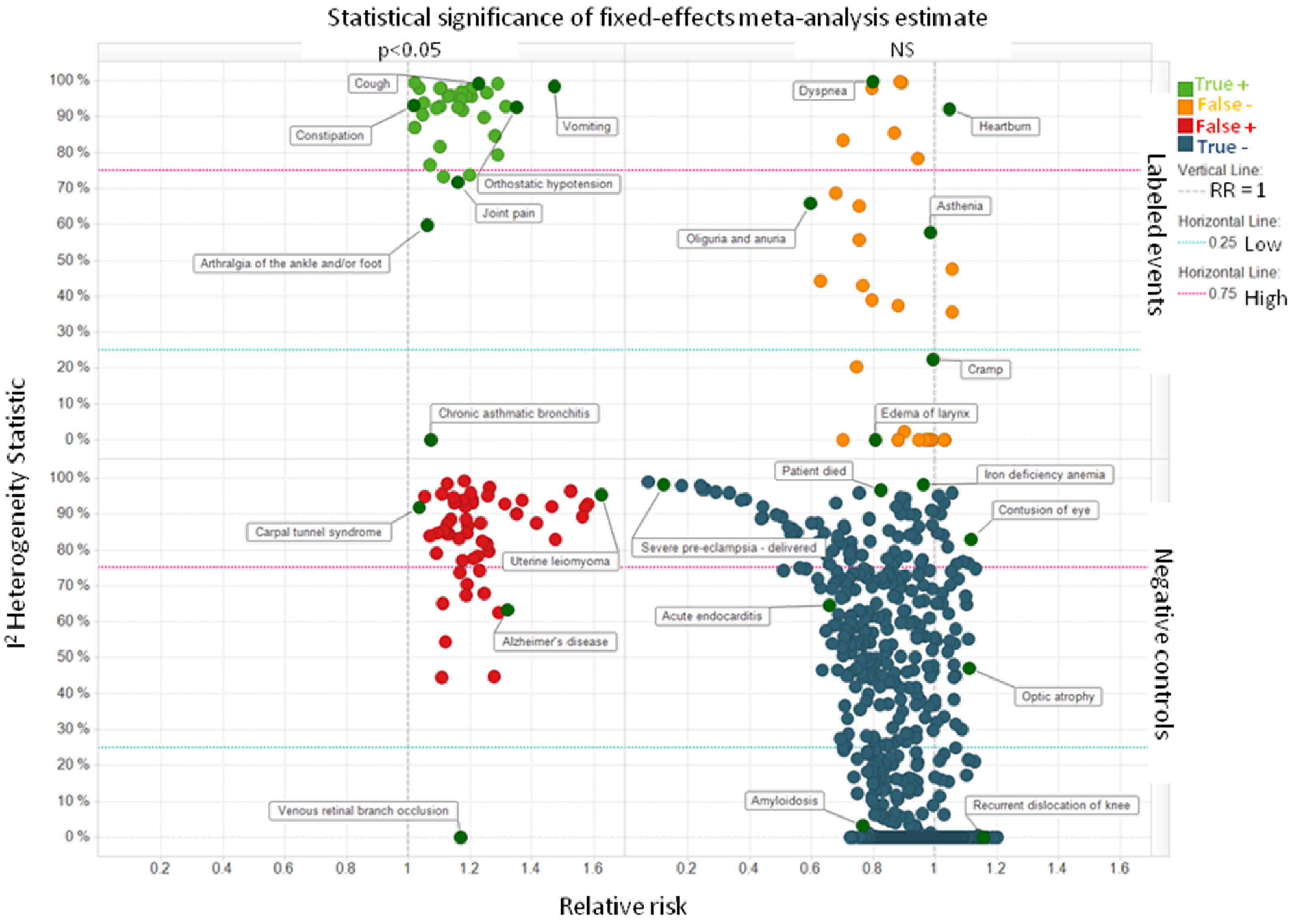


Figure 23: Heterogeneity across drug-outcome pair

## References

1. Public Law 110-85: Food and Drug Administration Amendments Act of 2007; 2007.

2. Platt R, Davis R, Finkelstein J, et al. Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics. *Pharmacoepidemiol Drug Saf.* Aug-Sep 2001;10(5):373-377.

3. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System - A National Resource for Evidence Development. *N Engl J Med.* Jan 12 2011.

4. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med.* Nov 2 2010;153(9):600-606.

5. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. *Pharmacoepidemiol Drug Saf.* 2010.

6. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology.* Jul 2009;20(4):512-522.

7. Ryan PB. Enhancing Drug Safety Through Active Surveillance of Observational Healthcare Data, Chapter 4: Systematic identification of drug safety issues in administrative claims data:  Performance of hypothesis generation methods for active surveillance. Chapel Hill, NC: Eshelman School of Pharmacy, University of North Carolina at Chapel Hill; 2011.

8. Furberg CD, Levin AA, Gross PA, Shapiro RS, Strom BL. The FDA and drug safety: a proposal for sweeping changes. *Arch Intern Med.* Oct 9 2006;166(18):1938-1942.

9. Ryan PB. *Defining a Reference Set for Evaluating the Performance of Active Surveillance Method*: Observational Medical Outcomes Partnership; 2011.

10. Chou R, Helfand M, Carson S. *Drug Class Review on Angiotensin Converting Enzyme Inhibitors* 1995.

11. Egger M, Smith GD, Altman D. *Systematic Reviews in Health Care: Meta-analysis in context*: BMJ Publishing Group; 2001.

12. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* Sep 6 2003;327(7414):557-560.

13. Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies. *BMJ.* Jan 10 1998;316(7125):140-144.

14. Egger M, Smith GD, Phillips AN. Meta-analysis: principles and procedures. *BMJ.* Dec 6 1997;315(7121):1533-1537.

15. Davey Smith G, Egger M, Phillips AN. Meta-analysis. Beyond the grand mean? *BMJ.* Dec 13 1997;315(7122):1610-1614.

16. Olkin I. Diagnostic statistical procedures in medical meta-analyses. *Stat Med.* Sep 15-30 1999;18(17-18):2331-2341.

17. Sterne JA, Egger M, Smith GD. Systematic reviews in health care: Investigating and dealing with publication and other biases in meta-analysis. *BMJ.* Jul 14 2001;323(7304):101-105.

18. Assimes TL, Elstein E, Langleben A, Suissa S. Long-term use of antihypertensive drugs and risk of cancer. *Pharmacoepidemiol Drug Saf.* Nov 2008;17(11):1039-1049.

19. Kedika R, Patel M, Pena Sahdala HN, Mahgoub A, Cipher D, Siddiqui AA. Long-term Use of Angiotensin Converting Enzyme Inhibitors Is Associated With Decreased Incidence of Advanced Adenomatous Colon Polyps. *J Clin Gastroenterol.* Feb 2011;45(2):e12-16.

20. Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol.* Oct 2001;54(10):1046-1055.

21. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ.* Sep 13 1997;315(7109):629-634.

22. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA.* Apr 19 2000;283(15):2008-2012.

# CHAPTER SIX: MANUSCRIPT 3:
# "Comparative safety of ACE inhibitors: Evaluating an active surveillance framework"

## Abstract

**Background:** Angiotensin-converting enzyme (ACE) inhibitors have proven effective treatments for hypertension. Product labeling indirectly suggests differences in adverse event profiles among ACE inhibitors, but little evidence exists about the comparative safety profile in real-world settings.

**Objectives:** To estimate and compare the risk of 23 adverse events among seven products within the ACE inhibitor class (lisinopril, benazepril, enalapril, ramipril, quinapril, captopril, and moexipril) by applying an active surveillance method against a large administrative claims database.

**Results:** Most risks were comparable across the ACE inhibitor class, though differential increased effects for ramipril (low blood pressure, RR=1.60 [95% CI 1.54-1.67]) and enalapril (orthostatic hypotension, RR=2.12 [95% CI 1.85-2.42]) were identified.

**Conclusions:** The safety profiles of products within the ACE inhibitor class are largely consistent, with differences in product labeling not observed in real-world study. Systematic use of observational databases for comparative safety assessment provides important, real-world evidence for decision-making.

**Key words:** Active surveillance; ACE inhibitors; Drug safety

## Introduction

Angiotensin-converting enzyme (ACE) inhibitors, along with diuretics, angiotensin II receptor blockers (ARBs), calcium channel blockers, and beta-blockers, offer providers and patients many options for pharmacologic treatment of hypertension. ACE inhibitors have been found to be effective in the control of blood pressure, reducing the risk of acute myocardial infarction among patients with heart failure, and decreasing progression of kidney damage among diabetic and hypertensive patients[1]. ACE inhibitors are generally well-tolerated, though are known to have potential side effects, such as cough, hyperkalemia, and hypotension, and in rare occasions, angioedema and renal dysfunction[2-7]. While there have been many placebo-controlled randomized trials and some head-to-head experiments synthesized in meta-analyses, little evidence to date has distinguished the efficacy or safety profile between the products within the class[1, 8-10]. Some observational studies have explored specific potential risks, such as congenital malformations[11], cancer[12, 13], and angioedema[14], but there have been no pharmacoepidemiologic studies examining the full comparative safety profile of ACE inhibitors.

There has been increasing interest in expanding the secondary use of large linked healthcare databases, such as administrative claims and electronic health records, towards the development of systems for active drug safety surveillance and comparative effectiveness research[15-18]. Such systems could apply standardized processes to identify and evaluate real-world effects of medicines, and to explore differentiated outcomes among alternative treatments. Several methodological issues exist to ensure the appropriate use of observational data for comparative studies[19-21], but few studies have been conducted that evaluate the performance of observational analyses toward this aim.

This study explores the comparative safety of seven products within the ACE inhibitor class (lisinopril, benazepril, enalapril, ramipril, quinapril, captopril, and moexipril) by applying an active surveillance method against a large administrative claims database. The aim of the study is to assess the incidence of side effects listed on the product labels, and to evaluate the potential utility of a standardized process for evidence generation.

## Methods

### Data Source

The study population used for this evaluation came from the Thomson Reuters MarketScan Commercial Claims and Encounters (CCAE), a large administrative claims database containing 59 million privately insured lives. CCAE provides patient-level de-identified data from inpatient and outpatient visits and pharmacy claims of multiple large employer-based health plans from 2003 to 2008. CCAE contains 3,052,264 persons with at least one prescription dispensing record for an ACE inhibitor, though the sample size available for each active ingredient varies by usage. The CCAE database was transformed into the Observational Medical Outcomes Partnership (OMOP) common data model[22], with all *International Classification of Diseases*, Ninth Revision (ICD-9) diagnosis codes translated into a standardized terminology using *Systematized Nomenclature of Medicine-Clinical Terms* (SNOMED CT) condition concepts.

### Product label review

Candidate outcomes were identified through natural language processing of all structured product labels within the class[23], and subsequently refined though manual review of product labels for each drug in the class[2-7]. A set of 23 outcomes were selected for exploration based on the adverse event being listed in the product label and the event being observable in claims data through ICD-9 diagnosis codes mapped in the OMOP standardized terminology[24].

**Statistical Analysis**

COMParator-Adjusted Safety Surveillance (COMPASS) is a statistical algorithm that estimates adjusted rate ratios for all outcomes of interest for a given medical product through propensity score stratification across exposed and unexposed cohorts within an incident user design. COMPASS applies an automated heuristic for defining a comparator group based on the indication of the medical product, and provides multivariate adjustment focused on minimizing bias from four primary sources: personal demographics (such as age and gender), confounding by indication, effects of comorbidity, and health service utilization.

COMPASS leverages large biomedical ontologies, or networks of clinical concepts such as relationships between diseases and treatments, to automate comparator selection by identifying all drugs that share at least one FDA-approved indication but have different mechanisms of action than the target drug of interest. For this study, the heuristic matched each product in the ACE inhibitor class to alternative treatments for the same indications (indications listed in Table 15). Comparator products that were included for all ACE inhibitors were: Angiotensin II receptor blockers, beta-blockers, calcium channel blockers, and diuretics. Ramipril's comparator group included statins and nitroglycerin. Captopril's comparator group also included nitroglycerin. Benazepril and enalapril were the only drugs to not include class IV antiarhythmics (diltiazem, verapamil) and amlodipine and felodipine (due to combination use). Digoxin was included in comparator groups for all drugs but benazepril and moexipril. Lisinopril and ramipril comparator groups included aspirin, clopidoprel, alteplase, and phenprocoumon.

COMPASS imposes automated study design heuristics, including cohort exclusion criteria based on contraindications and covariate selection based on FDA-approved indications and off-label uses. Persons included in this study were required to have a diagnosis code for at least one approved indication (in Table 1) any time prior to the index exposure, and were excluded if they had a diagnosis of a listed contraindication (such as pregnancy, liver disease, and renal artery stenosis, or prior angioedema, neutropenia, and hyperkalemia) within 30 days of exposure. Off-label uses, such as diastolic heart failure, prevention of recurrent atrial fibrillation, and renal crisis scleroderma, were used as covariates in the propensity score model. The number of prior medications dispensed for the indication was also used as a covariate to adjust for past treatment attempts. The Romano version of the Charlson comorbidity index and its constituent diseases[25, 26], as well as total conditions were used to adjust for disease burden[27]. Covariates for total prescriptions dispensed, total procedures, and total inpatient and outpatient visits 30 days prior to exposure were used as proxies to balance on health service utilization. Additional covariates included patient demographics (age at exposure and gender) and proxies for lifestyle risk factors: smoking, obesity, alcohol and drug abuse. The propensity score was estimated within each calendar year using multivariate logistic regression using all covariates described above to estimate probability of exposure classification, and the cohorts were stratified into 20 quantiles based on the propensity score distribution. Cohort balance pre- and post-adjustment was assessed using a heatmap visualization across all covariates and each products. Outcomes were identified as incident condition occurrences, based on date of recorded diagnosis codes, within 30 days from the date of first exposure for focus on acute onset events with close proximity to exposure. Outcome counts were tabulated within each propensity score strata, and adjusted relative risks were estimated through inverse variance weighting of the strata-specific effects. COMPASS was developed using SAS version 9.2 (Cary, NC).

COMPASS has previously been shown to have better performance than other viable active surveillance methods in identifying true drug safety issues and discerning from false positive findings when exploring ACE inhibitor class effects[28]. COMPASS has been studied across a network of five disparate databases, and CCAE was shown to have the highest sensitivity and best predictive model amongst the available sources[29].

Event counts and unadjusted incidence estimates were generated for each ACE inhibitor product and its corresponding active comparator group. Adjusted relative risks and 95% confidence intervals (CI) were calculated from COMPASS for each outcome.

**Results**

Table 15 shows the number of persons that were eligible for inclusion in the exposed and comparator cohorts within the inception cohort design. Lisinopril had the largest cohort size (n=339,556) while both captopril and moexipril had fewer than 10,000 persons meeting the study criteria. Each drug had a large sample used to define the comparator cohort, with at least 696,353 exposed to alternative treatment. The comparator cohort sizes vary due to differences in the FDA-approved indications for each treatment (highlighted in Table 15), as well as differing contraindications used as restriction criteria.

Each ingredient has notable differences in baseline characteristics among the exposed populations. Persons with incident ramipril exposure are older and more heavily weighted toward males than other products. Both ramipril and captopril cohorts have higher average disease severity, as measured by Charlson index, and higher health service utilization, in terms of total concomitant medications, procedures administered, and inpatient and outpatient visits prior to incident exposure. Ramipril and captopril patients, on average, had exposure to 2.3 prior medications for the respective indications, while patients with incident exposure to other ACE inhibitors had at least prior exposure to an alternative indicated treatment (1.35-1.87).

Figure 24 depicts the number of patients in the exposed and comparator cohorts within propensity score strata. For all products, there is substantial sample in the comparator group throughout all 20 strata; the 20th highest strata within lisinopril analysis offers the smallest sample (n=6906). However, there are substantial differences in the distribution of exposed persons across the propensity strata across products, indicating differential discrimination in the propensity score model. In the ramipril cohort, 65% of the exposures (n=32,771) fell within the top 5% of propensity scores. Captopril and moexipril have strata with fewer than 100 patients, while ramipril, quinapril, and enalapril have fewer than 1000 exposed in the lowest 5% propensity score strata.

Figure 25 depicts a heatmap that highlights the impact of propensity score adjustment on baseline characteristics, including patient demographics, health service utilization, comorbidities, lifestyle risk factors, and indication prevalence. Each column in the graph represents a covariate, and the color gradient within the column reflects the range of observed values across all cohorts (darker colors indicate higher mean values). Within each product, trellised in rows with ingredient name on the right, there are summary measures displayed for the exposed (1) and comparator (0) cohorts pre- and post-adjustment. Color differences between pre-0 and pre-1 rows reflect cohort imbalances in the mean of that attribute; for example, benazepril cohort had a higher proportion of males and greater number of prior indication medications than its comparator cohort (columns 2 and 3, rows 3 and 4). Balance after propensity score stratification can be observed by assessing post-1 and post-0 rows; the benazepril cohort was well-balanced with its comparator cohort on gender (column 2, rows 1 and 2) and indication medications column 3, rows 1 and 2), as evidenced by similar color. Ramipril and captopril exposed cohorts had the highest rates of comorbidities and risk factors (obesity, smoking, and alcohol abuse). In contrast, the moexipril cohort had the lowest rates of health service utilization and comorbidities, but the highest prevalence of prior hypertension diagnosis. For all products, imbalance between exposed and comparator groups was observed on multiple baseline characteristics, but the magnitude and directionality of those imbalances vary widely by product. Propensity score stratification achieves greater covariate balance for all products, but some residual imbalance is observed. For captopril, after adjustment the exposed group has higher prevalence of mild to moderate diabetes (0.17 vs. 0.11) and diabetic nephropathy (0.10 vs. 0.06), but lower prevalence of chronic pulmonary disorder (0.14 vs. 0.17) than the comparator cohort. For ramipril, the exposed cohort had 57% male vs. 51% in the comparator group, and higher prevalence of renal crisis scleroderma (0.46 vs. 0.38), but fewer incident exposures occurred within an inpatient visit (0.02 vs. 0.05).

Review of the product labels identified similarities and differences across the ACE inhibitors. Angioedema, hypotension, and leukopenia are listed in the warnings section of each label, while cough, hyperkalemia, and impaired renal function are cited as possible adverse events within the precautions section. Additional events consistently listed in the Adverse Reactions section of the product labels include: hemolytic anemia, constipation, diarrhea, dyspnea, end stage renal disease, abdominal pain, nausea, oliguria, palpitations, pruritis, and vomiting. Bronchospasm is not listed on the product labels for benazepril, quinapril, and ramipril. Asthma not listed for ramipril. Flushing is listed for all products, except quinapril and ramipril. Low back pain is not listed for captopril and ramipril. Tinnitus is on labels for all but captopril and quinapril. Table 16 shows the number of events and prevalence of these outcomes across the seven products. The active comparator for benazepril is provided as a benchmark, as it reflects the smallest comparator with the common indication across all products. Acquired hemolytic anemia, bronchospasm, edema of larynx, end stage renal disease, flushing, impaired renal function disorder, leucopenia, oliguria and anuria were observed to occur in fewer than 10 patients for all products, except lisinopril. Cough, dyspnea, low back pain and palpitations were the four most prevalent conditions across the cohorts. The unadjusted incidence of asthma amongst products with the event listed on the product label ranged from 0.60 to 3.00 per 1000 persons; ramipril (the only member of the class with asthma not listed on the label) had 98 cases for incidence of 1.94 / 1000 persons. Ramipril also did not have low back pain listed on the product label, but the unadjusted incidence of 4.53 was similar to those observed for all products with the labeled events. Two products, captopril and quinapril, that did not list tinnitus as an adverse event on the labels, had 3 and 4 cases, respectively.

Figure 26 shows the adjusted relative risks, estimated by COMPASS, for all products across 12 of the labeled events. Two relative risks were observed with lower bounds of 95%

confidence intervals > 1.5; low blood pressure for ramipril ( RR=1.60; 1.54-1.67) and orthostatic hypotension for enalapril (RR=2.12; 1.85-2.43).   Ramipril was the only product with statistically significant risks for asthma, dyspnea, hyperkalemia, and nausea.  Moexipril use was associated with significant lower risk of cough than comparator (RR=0.55; 0.49-0.60).  Low back pain for captopril appears lower than alternative treatments (RR=0.65; 0.56-0.75), while other ACE inhibitors show consistent relative risks with comparators. All other observed effects were small in magnitude and with concordance in direction among two or more products.

## Discussion

This is the first study to examine the full portfolio of potential side effects of ACE inhibitors in observational data through an active surveillance framework.  The observational analysis of real-world population complements the existing evidence from clinical trials, and provides a first side-by-side comparison of risks of individual ACE inhibitors relative to alternative treatment.  The analysis is derived from a large privately insured population, with over 535,000 new users of ACE inhibitors over a 5-year period.

The analysis suggests that the seven ACE inhibitors under study have largely comparable safety profiles, in terms of incidence rates of 23 events suggested from product labeling.  All but four events (asthma, cough, dyspnea, and palpitations) had unadjusted rates amongst ACE inhibitors that varied by less than 2 events per 1000 persons, and only dyspnea was observed within an ACE inhibitor (captopril) to have an incremental risk greater than 1 per 1000 persons relative to the active comparator.  However, the analysis also generated a few hypotheses of differential effects that may warrant further study.  Unlike the other drugs in the class, users of ramipril had a significantly higher incidence of low blood pressure than alternative treatments. Persons initiating enalapril had a two-fold increase in the risk of orthostatic hypotension. Captopril users were observed to have a lower incidence rate of low back pain than its comparator, whereas other ACE inhibitors had consistently similar rates to alternative treatments. Cough has been demonstrated in clinical trials to be one of the most prevalent adverse events for all ACE inhibitors. None of the ACE inhibitors were associated with clinically significant increased risks, and moexipril use was associated with significantly fewer cough events than its comparator.

Prior studies have suggested that the adverse event rate of ACE inhibitor-induced cough is far higher than those reported from clinical trials and in product labels[30-32]. Part of the difference between observational analyses can be attributed not only to the different source populations but also due to study design, as Visser et al.[30] conducted a case-control design to explore medication-related effects among cases of cough vs. matched controls, whereas COMPASS studied the relative effects of ACE inhibitors relative to other antihypertensives. Both studies may be susceptible to residual channeling bias, since providers are generally aware of the known side effect and may accordingly alter treatment decisions for those at risk[33]. Note also neither study performed source record verification to confirm the cough diagnoses.

A noteworthy finding from this study is the relative concordance of safety profiles of ACE inhibitors in light of observed differences in product labeling. The ramipril labeling does not report asthma, bronchospasm, flushing, or low back pain as adverse events on its US product label, but the observed rates within ramipril users appear consistent with other products in the class. Similarly, quinapril labeling does not include bronchospasm, flushing, or tinnitus, but no differences between quinapril and other ACE inhibitors were observed in this study. Only in one instance was the discrepancy between the labels consistent with differences observed in this study; the omission of low back pain on the captopril labeling is supported by the statistically significant decreased risk that was not observed with any other product. Perhaps most notable is that majority of the labeled events were not observed to occur at significantly different rates than alternative treatments.

Discrepancies between product labeling and real-world observation aren't necessarily unexpected. Product labels offer one important tool for product manufacturers and regulators to communicate with providers and patients about the potential effects of medicines. Serious side effects are often highlighted in boxed warnings, or described in the Warnings and Precautions sections of labels, while other adverse events are listed in the Adverse Reaction section[34]. The evidence used to support product labeling is generally based on randomized clinical trials, often using placebo as a comparator, prior to approval, although post-approval spontaneous adverse event reports may also be documented. Randomized trials are often limited by lack of generalizability and restricted sample sizes, such that the observed frequency of event occurrence can be small and reported differences between treatment arms may be statistically insignificant. Spontaneous adverse event reporting reflects case series of self-reported suspicions of drug-event relationships, but can suffer from substantial bias in underreporting and are limited in interpretation due to lack of denominator for contextualizing the rate of events relative to other drugs[35, 36]. In neither case is the intention to provide a relative assessment of comparative safety, but instead to provide an absolute reporting of adverse events that have been observed. FDA guidance suggest that labels should only include adverse event evidence that would be "useful to health care practitioners making treatment decisions"[37], based on frequency of occurrence or rates of discontinuation or suspicion of a causal effect. While product labels offer evidence of the occurrence of adverse events during the product lifecycle, they do not communicate the level of confidence in a causal attribution of the effect nor are they intended to communicate comparative differences between alternative treatments for the same indication.

Observational data offer the potential for providing more precise measures of risk using large samples, but observational analyses are susceptible to bias and confounding. Pharmacoepidemiologic evaluation studies are increasingly becoming an important source of post-approval evidence, but typically focus on pre-specified hypotheses of risk and are often not conducted in a consistent and reproducible fashion. Observational database networks for active surveillance and comparative effectiveness can enable a systematic process for evidence generation that can provide ongoing assessment of the both the beneficial and negative effects of medicines, relative to alternative treatments, in real-world populations. However, as these efforts continue their development, it raises the need for a central evidence source that comprise all existing information, from pre-clinical studies, clinical trials, spontaneous reporting, and observational data, that can be used to support safety and effectiveness assessments and inform medical decision making about alternative treatments. Currently, evidence like that produced in this study, does not have a logical home beyond the peer-review literature.

It is important to reinforce that COMPASS is intended as a hypothesis generating tool, used within an active surveillance system as a standardized process to enable proactive monitoring of medical products to identify and refine hypotheses about potential drug effects. While COMPASS applies many of the analytical and design choices typically considered within a pharmacoepidemiologic study[20, 38-40], the results should be considered exploratory in nature and interpreted with caution accordingly. In particular, researchers should consider additional sources of confounding that may have been inadequately addressed in the COMPASS model. For example, while COMPASS uses covariates of lifestyle risk factors, such as smoking, obesity, tobacco and alcohol use, these variables are known to be poorly represented in administrative claims data and are not an accurate reflection of these characteristics in the population[41]. There may also be a need to examine the validity of the outcomes, which are initially defined only by diagnosis codes without source record verification. The analysis should be embedded within a broader sensitivity analysis framework whereby alternative decisions for issues like time-at-risk definition, covariate selection, and adjustment strategy can be systematically evaluated. Each potential association uncovered in this exploratory phase of active surveillance requires further evaluation as part of a causal assessment. Other evidence to raise confidence in the potential effect beyond the temporal association that is observed in these observational databases is the biological plausibility of the relationship based on clinical pharmacology and additional evidence of a dose-response relationship.

Here, COMPASS has been used to evaluate the comparative safety of products within a specific class through indirect comparisons. These indirect comparisons complicate the interpretation of results, but are necessary because each product has different indications and contraindications that should be accounted for in the construction of a proper referent group with similar baseline characteristics. Each product was evaluated against a proxy for standard of care by comparing effects patients exposed to the product with patients exposed to alternative treatments for the same indications as the target drug. In the case of ACE inhibitors, all products share a common FDA-approved indication for hypertension, but some products have approved indications for other conditions, such as chronic heart failure, prevention of myocardial infarction and stroke, and diabetic nephropathy. The impact of adjustment via propensity score is different across products since the underlying populations reflect different patient demographics, health service utilization patterns, comorbidities and other risk factors. The observed differences underscore the importance of a thorough examination of comparator populations prior to assessment of outcome differences[39]. In this instance, the substantial heterogeneity in baseline characteristics between products in the ACE inhibitor class would have been obscured had the drugs not been analyzed separately. Further methodological research is needed to establish best practices for integrating and evaluating evidence through indirect comparisons within therapeutic classes and across observational databases. Alternative analytical approaches should also be evaluated and compared with COMPASS. This may include strategies to enable direct comparisons across the ACE inhibitor products, with further adjustment to address the different indicated populations and concerns for channeling bias.

This study serves as a proof-of-concept to demonstrate the opportunities and challenges awaiting the development of a national active surveillance system. While the safety profile of ACE inhibitors is generally thought to be well-established, COMPASS provided corroborating evidence about the magnitude of risk that is consistent across the drug class, highlighted substantial differences in treatment patterns among the products, and identified hypotheses about potential differential effects that may warrant further evaluation. It bears mentioning that there is currently little incentive for comprehensive research on mature products that are off-patent, such as ACE inhibitors. However, given their widespread use and cost-effectiveness, it could be argued that relative assessments of mature products could provide the most impactful evidence to improve patient health while reducing overall healthcare costs. Further work will be needed to establish the operating characteristics of COMPASS and other potential active surveillance methods across a broader array of medical products to understand the reliability of the evidence the system can provide. Once established, it could be anticipated such as system could directly support the assessment of newly marketed medicines with emerging safety concerns. More broadly, the availability of comparative evidence of medical therapies should support all stakeholders in the healthcare community, including product manufacturers, regulators, payors, healthcare systems, and clinicians, in maximizing the care provided to patients.

## Tables and Figures

Table 15: Cohorts, baseline characteristics, indications

| | Lisinopril | Benazepril | Enalapril | Ramipril | Quinapril | Captopril | Moexipril |
|---|---|---|---|---|---|---|---|
| Incident users of drug | 339,556 | 96,325 | 27,750 | 50,560 | 12,900 | 3,330 | 5,024 |
| Comparator cohort | 699,946 | 696,353 | 703,574 | 1,126,697 | 778,472 | 828,155 | 794,778 |
| Baseline characteristics | | | | | | | |
| Age (mean) | 50.75 | 50.77 | 50.30 | 53.36 | 51.18 | 51.75 | 50.27 |
| Gender (% males) | 0.53 | 0.55 | 0.52 | 0.58 | 0.53 | 0.48 | 0.46 |
| Charlson index (mean) | 0.76 | 0.68 | 0.86 | 1.36 | 0.79 | 1.25 | 0.56 |
| Indication medication count | 1.55 | 1.87 | 1.60 | 2.38 | 1.71 | 2.35 | 1.35 |
| Total medication count | 0.75 | 0.78 | 0.77 | 1.09 | 0.78 | 0.93 | 0.76 |
| Total procedure count | 1.67 | 1.57 | 1.57 | 4.28 | 1.52 | 2.58 | 1.50 |
| Total outpatient visits | 1.17 | 1.26 | 1.16 | 2.00 | 1.19 | 1.36 | 1.28 |
| Total inpatient visits | 0.09 | 0.06 | 0.12 | 0.29 | 0.08 | 0.28 | 0.05 |
| FDA-approved indications | | | | | | | |
| Asymptomatic Left Ventricular Dysfunction | | | x | | | | |
| Chronic Heart Failure | x | | x | x | x | x | |
| Diabetic Nephropathy | | | | | | x | |
| Hypertension | x | x | x | x | x | x | x |
| Left Ventricular Dysfunction following Myocardial Infarction | | | | x | | x | |
| Myocardial Infarction | x | | | | | | |
| Myocardial Infarction Prevention | | | | x | | | |
| Prevention of Cerebrovascular Accident | | | | x | | | |

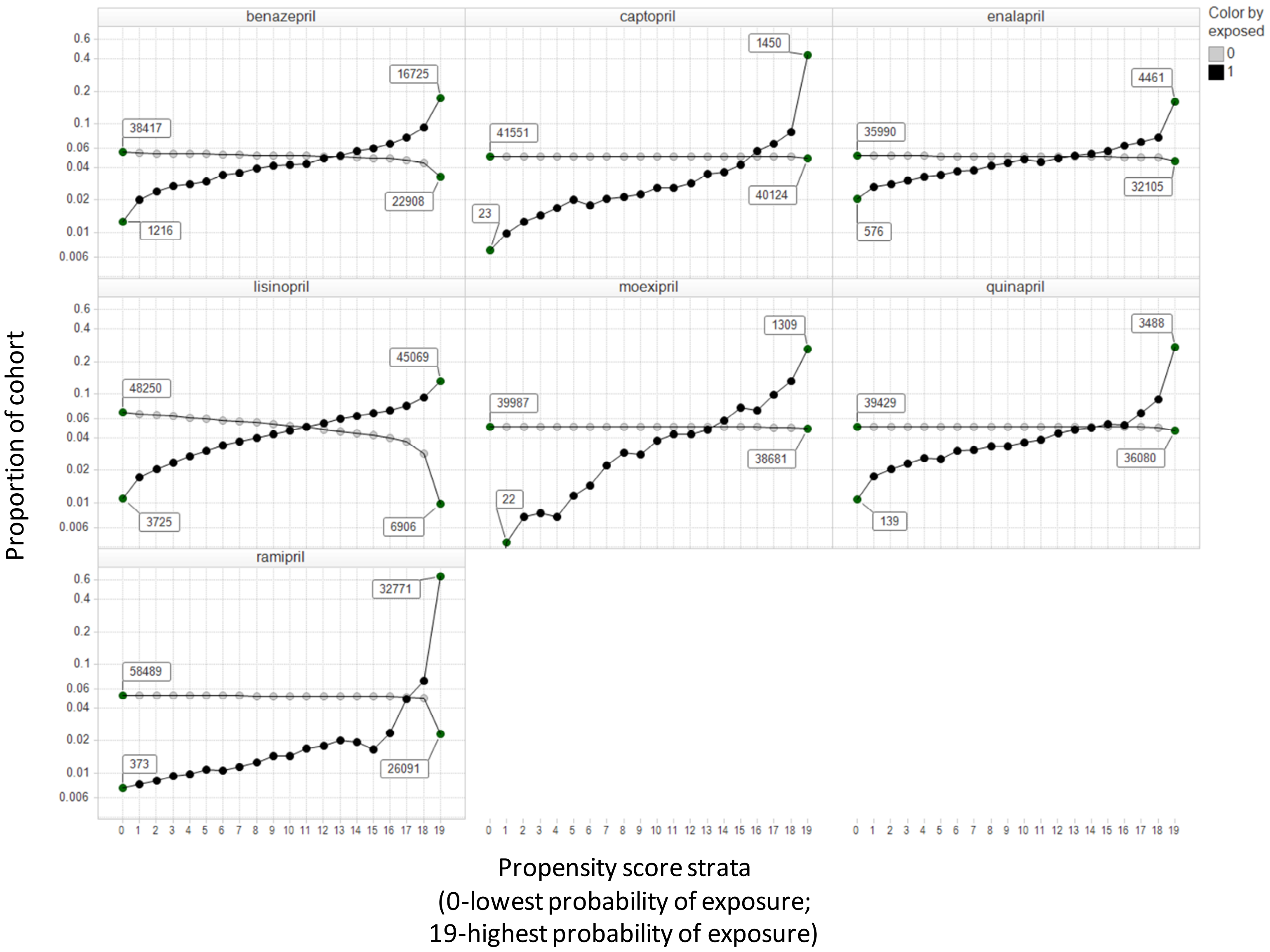


Figure 24: Population size by propensity score strata

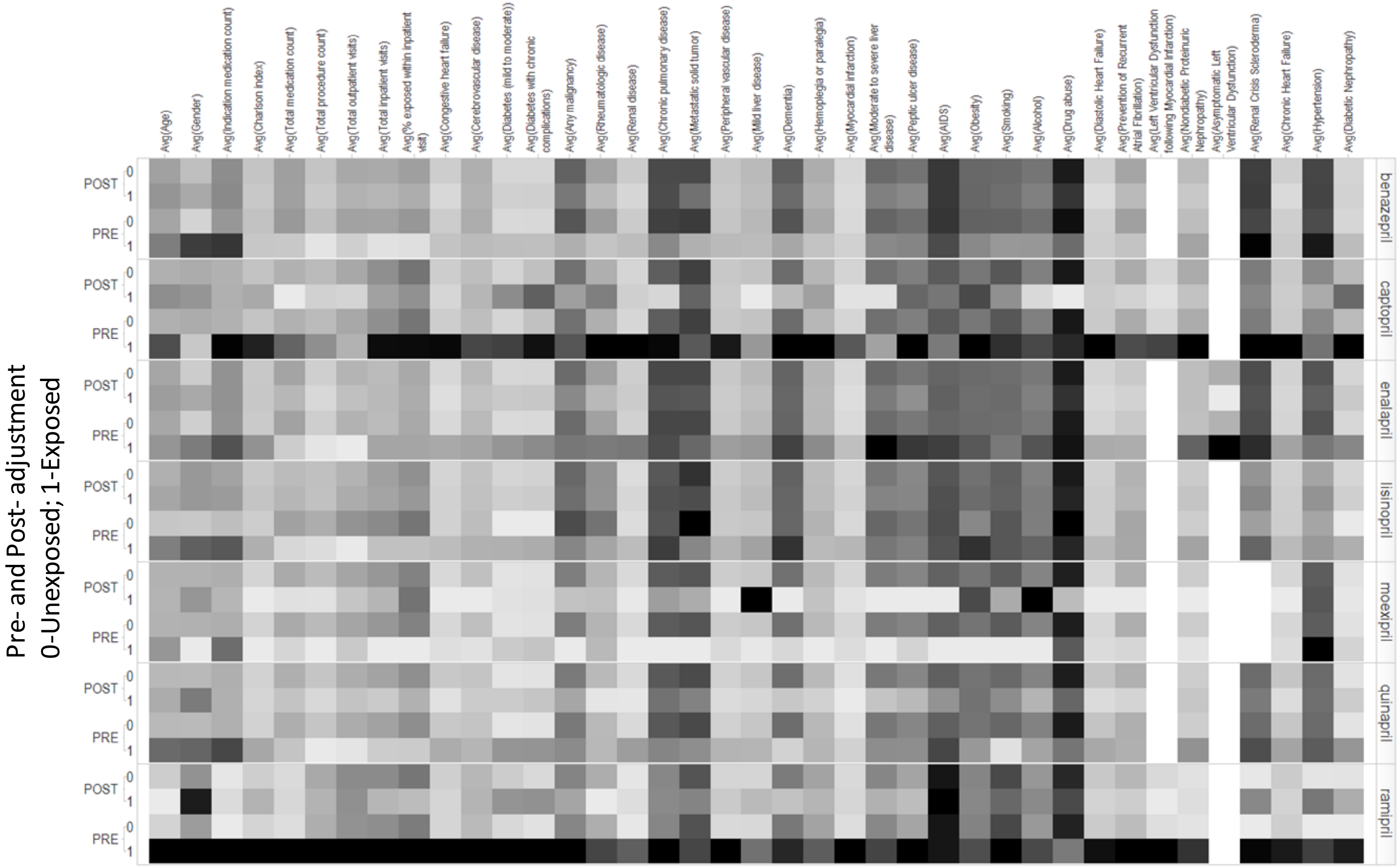


Figure 25: Impact of propensity score adjustment on measured covariates

Table 16: Event rates by ingredient

|  | Unexposed comparator |  | Benazepril |  | Captopril |  | Enalapril |  | Lisinopril |  | Moexipril |  | Quinapril |  | Ramipril |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ACE Inhibitor Labeled Events (exceptions bold in grey) | Events | Events/ 1000 persons | Events | Events/ 1000 persons | Events | Events/ 1000 persons | Events | Events/ 1000 persons | Events | Events/ 1000 persons | Events | Events/ 1000 persons | Events | Events/ 1000 persons | Events | Events/ 1000 persons |
| Acquired hemolytic anemia | 4 | 0.01 | 1 | 0.01 | 0 | 0.00 | 0 | 0.00 | 4 | 0.01 | 0 | 0.00 | 0 | 0.00 | 1 | 0.02 |
| Asthma | 1534 | 2.20 | 143 | 1.48 | 10 | 3.00 | 57 | 2.05 | 593 | 1.75 | 3 | 0.60 | 24 | 1.86 | **98** | **1.94** |
| Bronchospasm | 84 | 0.12 | **3** | **0.03** | 0 | 0.00 | 2 | 0.07 | 40 | 0.12 | 1 | 0.20 | **2** | **0.16** | **2** | **0.04** |
| Constipation | 970 | 1.39 | 120 | 1.25 | 4 | 1.20 | 47 | 1.69 | 404 | 1.19 | 8 | 1.59 | 15 | 1.16 | 64 | 1.27 |
| Cough | 3713 | 5.33 | 486 | 5.05 | 21 | 6.31 | 168 | 6.05 | 1892 | 5.57 | 19 | 3.78 | 67 | 5.19 | 275 | 5.44 |
| Diarrhea | 1390 | 2.00 | 137 | 1.42 | 10 | 3.00 | 51 | 1.84 | 682 | 2.01 | 8 | 1.59 | 16 | 1.24 | 101 | 2.00 |
| Dyspnea | 6184 | 8.88 | 460 | 4.78 | 33 | 9.91 | 181 | 6.52 | 2003 | 5.90 | 16 | 3.18 | 69 | 5.35 | 475 | 9.39 |
| Edema of larynx | 19 | 0.03 | 1 | 0.01 | 0 | 0.00 | 0 | 0.00 | 14 | 0.04 | 1 | 0.20 | 0 | 0.00 | 2 | 0.04 |
| End stage renal disease | 89 | 0.13 | 5 | 0.05 | 0 | 0.00 | 3 | 0.11 | 33 | 0.10 | 0 | 0.00 | 1 | 0.08 | 4 | 0.08 |
| Flushing | 99 | 0.14 | 9 | 0.09 | 0 | 0.00 | 1 | 0.04 | 22 | 0.06 | 0 | 0.00 | **3** | **0.23** | **3** | **0.06** |
| Generalized abdominal pain | 1370 | 1.97 | 137 | 1.42 | 6 | 1.80 | 59 | 2.13 | 598 | 1.76 | 7 | 1.39 | 24 | 1.86 | 100 | 1.98 |
| Hyperkalemia | 492 | 0.71 | 44 | 0.46 | 5 | 1.50 | 19 | 0.68 | 208 | 0.61 | 1 | 0.20 | 6 | 0.47 | 49 | 0.97 |
| Impaired renal function disorder | 49 | 0.07 | 7 | 0.07 | 1 | 0.30 | 2 | 0.07 | 31 | 0.09 | 0 | 0.00 | 1 | 0.08 | 5 | 0.10 |
| Leukopenia | 40 | 0.06 | 3 | 0.03 | 1 | 0.30 | 1 | 0.04 | 19 | 0.06 | 0 | 0.00 | 1 | 0.08 | 2 | 0.04 |
| Low back pain | 3226 | 4.63 | 425 | 4.41 | **13** | **3.90** | 128 | 4.61 | 1578 | 4.65 | 22 | 4.38 | 57 | 4.42 | **229** | **4.53** |
| Low blood pressure | 623 | 0.89 | 36 | 0.37 | 3 | 0.90 | 18 | 0.65 | 282 | 0.83 | 4 | 0.80 | 8 | 0.62 | 65 | 1.29 |
| Nausea | 942 | 1.35 | 86 | 0.89 | 6 | 1.80 | 30 | 1.08 | 397 | 1.17 | 7 | 1.39 | 8 | 0.62 | 65 | 1.29 |
| Oliguria and anuria | 18 | 0.03 | 1 | 0.01 | 0 | 0.00 | 0 | 0.00 | 12 | 0.04 | 0 | 0.00 | 1 | 0.08 | 0 | 0.00 |
| Orthostatic hypotension | 205 | 0.29 | 17 | 0.18 | 1 | 0.30 | 16 | 0.58 | 109 | 0.32 | 1 | 0.20 | 3 | 0.23 | 17 | 0.34 |
| Palpitations | 4625 | 6.64 | 251 | 2.61 | 12 | 3.60 | 82 | 2.95 | 1155 | 3.40 | 9 | 1.79 | 49 | 3.80 | 248 | 4.91 |
| Pruritus of skin | 181 | 0.26 | 22 | 0.23 | 2 | 0.60 | 8 | 0.29 | 83 | 0.24 | 1 | 0.20 | 4 | 0.31 | 13 | 0.26 |
| Tinnitus | 326 | 0.47 | 44 | 0.46 | **3** | **0.90** | 14 | 0.50 | 139 | 0.41 | 0 | 0.00 | **4** | **0.31** | 19 | 0.38 |
| Vomiting | 473 | 0.68 | 52 | 0.54 | 5 | 1.50 | 22 | 0.79 | 241 | 0.71 | 4 | 0.80 | 5 | 0.39 | 27 | 0.53 |

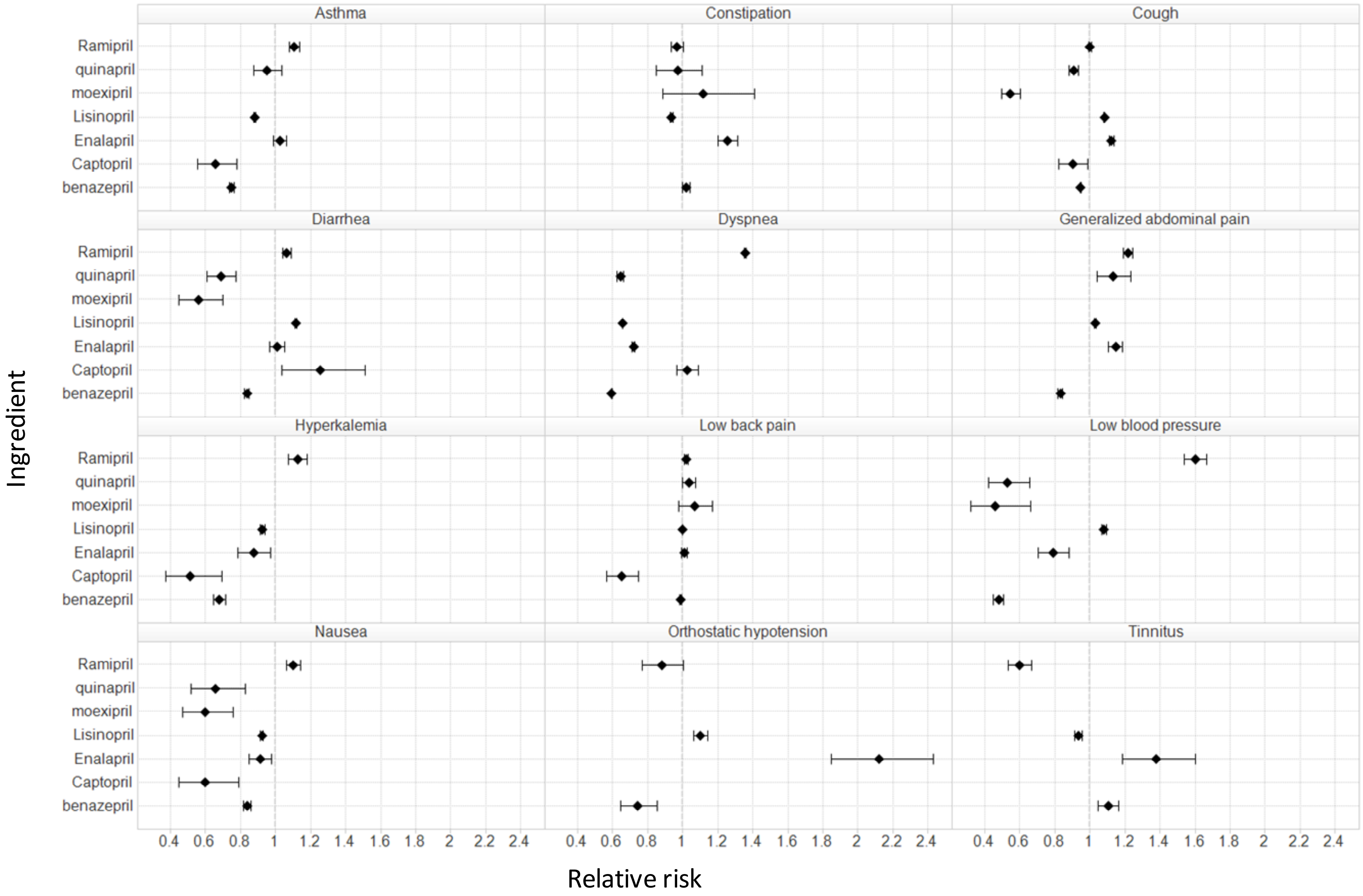


Figure 26: Effects estimates by outcome

## References

1. Norris, S., et al., *Drug class review: Direct renin inhibitors, angiotensin converting enzyme inhibitors, and angiotensin II receptor blockers.* 2010.

2. FDA. *Zestril (lisinopril)*. [cited Feb 8, 2011]; Available from: http://www.accessdata.fda.gov/drugsatfda_docs/label/2009/019777s056lbl.pdf.

3. FDA. *Lotensin (benazepril)*. [cited Feb 8, 2011]; Available from: http://www.accessdata.fda.gov/drugsatfda_docs/label/2009/019851s038lbl.pdf.

4. FDA. *Vasotec (enalapril maleate)*. [cited Feb 8, 2011]; Available from: http://www.accessdata.fda.gov/drugsatfda_docs/label/2008/018998s071lbl.pdf.

5. FDA. *Altace (ramipril)*. [cited Feb 8, 2011]; Available from: http://www.accessdata.fda.gov/drugsatfda_docs/label/2010/019901s055lbl.pdf.

6. FDA. *Accupril (quinapril)*. [cited Feb 8, 2011]; Available from: http://www.accessdata.fda.gov/drugsatfda_docs/label/2009/019885s028lbl.pdf.

7. FDA. *Univasc (moexipril)*. [cited Feb 8, 2011]; Available from: http://www.accessdata.fda.gov/drugsatfda_docs/label/2010/020312s031lbl.pdf.

8. Heran, B.S., et al., *Blood pressure lowering efficacy of angiotensin converting enzyme (ACE) inhibitors for primary hypertension.* Cochrane Database Syst Rev, 2008(4): p. CD003823.

9. Matchar, D.B., et al., *Systematic review: comparative effectiveness of angiotensin-converting enzyme inhibitors and angiotensin II receptor blockers for treating essential hypertension.* Ann Intern Med, 2008. 148(1): p. 16-29.

10. Chou, R., M. Helfand, and S. Carson, *Drug Class Review on Angiotensin Converting Enzyme Inhibitors*. 1995.

11. Cooper, W.O., et al., *Major congenital malformations after first-trimester exposure to ACE inhibitors.* N Engl J Med, 2006. 354(23): p. 2443-51.

12. Fryzek, J.P., et al., *A cohort study of antihypertensive medication use and breast cancer among Danish women.* Breast Cancer Res Treat, 2006. 97(3): p. 231-6.

13. Sjoberg, T., L.A. Garcia Rodriguez, and M. Lindblad, *Angiotensin-converting enzyme inhibitors and risk of esophageal and gastric cancer: a nested case-control study.* Clin Gastroenterol Hepatol, 2007. 5(10): p. 1160-1166 e1.

14. Miller, D.R., et al., *Angioedema incidence in US veterans initiating angiotensin-converting enzyme inhibitors.* Hypertension, 2008. 51(6): p. 1624-30.

15. Behrman, R.E., et al., *Developing the Sentinel System - A National Resource for Evidence Development.* N Engl J Med, 2011.

16. Platt, R., et al., *The new Sentinel Network--improving the evidence of medical-product safety.* N Engl J Med, 2009. 361(7): p. 645-7.

17. Maro, J.C., et al., *Design of a National Distributed Health Data Network.* Ann Intern Med, 2009.

18. Holmes, J.H., et al., *Developing a distributed research network to conduct population-based studies and safety surveillance.* AMIA Annu Symp Proc, 2008: p. 973.

19. Strom, B.L., *Methodologic challenges to studying patient safety and comparative effectiveness.* Med Care, 2007. 45(10 Supl 2): p. S13-5.

20. Schneeweiss, S., *A basic study design for expedited safety signal evaluation based on electronic healthcare data.* Pharmacoepidemiol Drug Saf, 2010.

21. Schneeweiss, S., *Developments in post-marketing comparative effectiveness research.* Clin Pharmacol Ther, 2007. 82(2): p. 143-56.

22. Ryan, P.B., et al. *OMOP Common Data Model (CDM) Specifications*. 2009 [cited 30 May 2010]; Available from: http://omop.fnih.org/CDMandTerminologies.

23. Ryan, P.B. *Defining a Reference Set for Evaluating the Performance of Active Surveillance Method*. 2010 [cited 2011 January 3, 2011]; Available from: http://omop.fnih.org/OMOPWhitePapers.

24. Reich, C. *OMOP Standard Vocabulary Specifications*. 2009 [cited November 28, 2009]; Available from: http://omop.fnih.org/CDMandTerminologies.

25. D'Hoore, W., A. Bouckaert, and C. Tilquin, *Practical considerations on the use of the Charlson comorbidity index with administrative data bases.* J Clin Epidemiol, 1996. 49(12): p. 1429-33.

26. Charlson, M., et al., *Validation of a combined comorbidity index.* J Clin Epidemiol, 1994. 47(11): p. 1245-51.

27. Schneeweiss, S. and M. Maclure, *Use of comorbidity scores for control of confounding in studies using administrative databases.* Int J Epidemiol, 2000. 29(5): p. 891-8.

28. Ryan PB. Enhancing Drug Safety Through Active Surveillance of Observational Healthcare Data, Chapter 4: Systematic identification of drug safety issues in administrative claims data: Performance of hypothesis generation methods for active surveillance. Chapel Hill, NC: Eshelman School of Pharmacy, University of North Carolina at Chapel Hill; 2011.

29. Ryan PB. Enhancing Drug Safety Through Active Surveillance of Observational Healthcare Data, Chapter 5: Integrating active drug safety surveillance analyses across a network of observational healthcare databases. Chapel Hill, NC: Eshelman School of Pharmacy, University of North Carolina at Chapel Hill; 2011.

30. Visser, L.E., et al., *Cough due to ACE inhibitors: a case-control study using automated general practice data.* Eur J Clin Pharmacol, 1996. 49(6): p. 439-44.

31. Visser, L.E., et al., *Angiotensin converting enzyme inhibitor associated cough: a population-based case-control study.* J Clin Epidemiol, 1995. 48(6): p. 851-7.

32. Bangalore, S., S. Kumar, and F.H. Messerli, *Angiotensin-converting enzyme inhibitor associated cough: deceptive information from the Physicians' Desk Reference.* Am J Med. 123(11): p. 1016-30.

33. Mackay, F.J., G.L. Pearce, and R.D. Mann, *Cough and angiotensin II receptor antagonists: cause or confounding?* Br J Clin Pharmacol, 1999. 47(1): p. 111-4.

34. FDA, *Guidance for Industry - Warnings and Precautions, Contraindications, and Boxed Warning Sections of Labeling for Human Prescription Drug and Biological Products — Content and Format* 2006.

35. Egberts, T.C., *Signal detection: historical background.* Drug Saf, 2007. 30(7): p. 607-9.

36. Meyboom, R.H., et al., *Pharmacovigilance in perspective.* Drug Saf, 1999. 21(6): p. 429-47.

37. FDA, *Guidance for Industry- Adverse Reactions Section of Labeling for Human Prescription Drug and Biological Products — Content and Format*. 2006.

38. Rubin, D.B., *On principles for modeling propensity scores in medical research.* Pharmacoepidemiol Drug Saf, 2004. 13(12): p. 855-7.

39. Rubin, D.B., *Estimating causal effects from large data sets using propensity scores.* Ann Intern Med, 1997. 127(8 Pt 2): p. 757-63.

40. Schneeweiss, S., et al., *Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results.* Med Care, 2007. 45(10 Supl 2): p. S131-42.

41. Eng, P.M., et al., *Supplementary data collection with case-cohort analysis to address potential confounding in a cohort study of thromboembolism in oral contraceptive initiators matched on claims-based propensity scores.* Pharmacoepidemiol Drug Saf, 2008. 17(3): p. 297-305.

# CHAPTER SEVEN: CONCLUSION AND DISCUSSION

## 7.1. Motivation for study

In the recent past, three remarkable forces have come together that have substantially raised the importance of ‘real-world’ data in understanding key outcomes of health care: advances in health IT, regulatory imperatives, and public / political activism in assuring the safety of medications.

The secondary use of automated healthcare databases, such as administrative claims and electronic health records, has been a cornerstone in pharmacoepidemiology, health outcomes and services research for many years. Researchers with access to these data sources have designed observational studies to examine a safety issue reported to be associated with a medical product, to compare alternative therapies for a given disease, and to assess the impact of new interventions within the healthcare delivery system on health service utilization and quality of care. While it has long been recognized that observational studies can suffer from various biases not present in an experimental setting, observational analyses remain a particularly valuable component of the evidence generation process for healthcare. In settings where prospective randomized trials are infeasible or unethical, such as the study of rare safety events with a latent onset following intervention, observational results may be the best evidence available to inform medical decision-making.

Recent advances in health information technology have increased capture of observational healthcare data and raised interest in coordinating large-scale efforts to leverage these data to better understand the effects of medical treatments. A recent report from the President's Council of Advisors on Science and Technology highlights the opportunities for how "improved health IT can directly affect, and improve, clinical encounters between doctor and patient, healthcare organizations, clinical research, and the monitoring of public health.[242]"

In the US, several national efforts offer the promise to significantly expand the use of observational data for evidence development. In 2007, Congress passed the Food and Drug Administration (FDA) Amendment Act, which called for the establishment of an "active postmarket risk identification and analysis system" with access to data from 100 million lives by 2012[11]. It is envisioned that an active surveillance system would "use sophisticated statistical methods to actively search for patterns in prescription, outpatient, and inpatient data systems that might suggest the occurrence of an adverse event, or safety signal, related to drug therapy"[243].

This reflects a significant evolution in the use of these data from the customized design of an individual study of a particular drug-outcome association applied to specific database at single point in time to the development of a systematic solution to a broader effort that effectively uses these data for active monitoring of any medical product and any health outcome of interest across a network of disparate databases. The envisioned system would go beyond the retrospective evaluation of hypothesized effects to proactively explore the data to generate and refine hypotheses of potential issues that warrant further scrutiny.

In January 2011, as part of its Sentinel Initiative, FDA announced it had the "capacity to 'query' the electronic health information of more than 60 million people, posing specific questions in order to monitor the safety of approved medical products"[244]. This initial focus on traditional pharmacoepidemiology evaluation studies of 'specific questions' supports the notion held by some that further research is needed to establish appropriate methods and gain understanding of operating characteristics prior to the system's more widespread use. Consistent with the trends in networks of data sources and the investigation of new methods, The Observational Medical Outcomes Partnership (OMOP), a public-private partnership chaired by the FDA and managed through the Foundation for the National Institutes of Health, is conducting methodological research to inform these national efforts by empirically measuring the performance of an array of alternative analysis methods across a network of 10 databases covering over 200 million patient lives[245]. Similar efforts are underway in Europe to assess performance of active surveillance methods across international data sources, including IMI-PROTECT[246] and EU-ADR[247].

Within the American Recovery and Reinvestment Act of 2009, $1.1 billion was committed to comparative effectiveness research (CER). The Federal Coordinating Council for Comparative Effectiveness Research defines CER as "the conduct and synthesis of research comparing the benefits and harms of different interventions and strategies to prevent, diagnose, treat and monitor health conditions in "real world" settings. The purpose of this research is to improve health outcomes by developing and disseminating evidence-based information to patients, clinicians, and other decision-makers, responding to their expressed needs about which interventions are most effective for which patients under specific circumstances."[248] A key priority within this investment is establishing a data infrastructure that provides coordinated linkage and access to administrative claims and electronic health records to enable research of medical interventions by a broad array of stakeholders.

While the opportunities abound, substantial research is needed to inform the appropriate use of observational healthcare data for active drug safety surveillance and comparative effectiveness research. Empirical studies are needed to determine the contribution of individual data sources into an observational data network and gain understanding of the performance characteristics of analytical methods when applied across the network in their ability to provide reliable evidence about the effects of medical products. This dissertation provides one body of research examining the use of a novel method across a network of observational databases to study the comparative safety of Angiotensin Converting Enzyme (ACE) Inhibitors.

### 7.2. Review of study results

This dissertation compiles a series of efforts intended to shed some light on the potential opportunities and challenges of an active surveillance system. First, it introduces a new method, COMPASS, designed to integrate standard pharmacoepidemiology principles into a systematic process for drug-outcome risk identification. The method was then applied in these experimental contexts to evaluate its performance, relative to other existing methods, and across a network of disparate observational databases. Finally, the method was applied to the specific clinical context of the comparative safety of ACE inhibitors to assess its potential utility as a tool for evidence generation.

COMParator-Adjusted Safety Surveillance (COMPASS) is a statistical algorithm that estimates adjusted rate ratios for all outcomes of interest for a given medical product through propensity score stratification across exposed and unexposed cohorts within an incident user design. COMPASS applies an automated heuristic for defining a comparator group based on the indication of the medical product and provides multivariate adjustment focused on minimizing bias from four primary sources: personal demographics (such as age and gender), confounding by indication, effects of comorbidity, and health service utilization. COMPASS was developed as a systematic process to support active surveillance, designed to incorporate basic epidemiologic principles typically used for evaluation studies of specific drug-outcome hypotheses but adapted to enable efficient, scalable analyses for proactive monitoring of multiple products and multiple outcomes simultaneously. As such, COMPASS applies a consistent set of heuristics within the framework to approximate the subjective decisions typically made during an evaluation design, such as comparator selection, inclusion/exclusion criteria, covariate adjustment strategy, and time-at-risk definition. It is important to reinforce that COMPASS is designed as an automated surveillance tool intended to supplement, not replace, existing pharmacovigilance practice. The outstanding question this research sought to address is whether COMPASS can provide useful supplemental information, as compared to other active surveillance methods under consideration.

In "Systematic identification of drug safety issues in administrative claims data: Performance of hypothesis generation methods for active surveillance," the first aim was addressed by characterizing the performance of COMPASS in identifying known safety issues association with ACE inhibitor exposure within an administrative claims database.

This study compared the operating characteristics of COMPASS with three existing active surveillance methods (disproportionality analysis, observational screening, and self-controlled case series) within the Thomson MarketScan Commercial Claims and Encounters database. Each method was applied to ACE inhibitor exposure, studying the same set of potential adverse events to assess the method's discrimination between true positives (events listed on the product labeled which are known to be associated with ACE inhibitors, such as cough, hypotension, and renal dysfunction) and negative controls (events not believed to be drug-related).

Amongst the four methods, COMPASS generated the fewest safety signals (statistically significant positive associations), had the lowest false positive rate, highest predictive probability and greatest precision. COMPASS was the only method to have specificity > 0.95. Given that COMPASS employed the most comprehensive strategy for addressing potential bias from between-person comparisons, it is reasonable to suggest that many of the false positives identified by disproportionality analysis and observational screening could be successfully mitigated through confounding adjustment. While COMPASS has the highest discrimination of the four methods (AUC=0.648), the absolute performance demonstrates the significant opportunity for method improvement. While COMPASS, along with all methods, performs substantially better than random, they are far from perfectly predictive models. The observed sensitivity of 0.42 for COMPASS suggests further work is needed to ensure that the strategies applied are not too restrictive as to fail to identify true relationships.

COMPASS's positive predictive value (0.31), while substantially better than the other three methods, underscores the risk of a active surveillance system to generate a majority of hypotheses that are false positives. The tolerance of false positives comes at a tradeoff for acceptability of false negatives. Only through empirical studies that provide experimental evidence of the operating characteristics of the active surveillance methods can stakeholders begin to establish policies for interpreting surveillance results. As a frame of reference, many have held the Vaccine Safety Datalink (VSD) as the bellwether for successful implementation of an active surveillance system, as it has been used to enable the study of newly marketed vaccines across a network of health maintenance organization claims databases. A recent assessment of the performance of the VSD system- which applies an unadjusted cohort design within the maximized sequential probability ratio testing framework- has highlighted that 9 of the 10 signals generated were subsequently determined to be false positives[249], which would be the equivalent of PPV=0.10. Given that epidemiologic study of vaccine exposure is often less susceptible to challenges with confounding (since most vaccines are administered to a more homogenous healthy infant or adolescent population), the fact that COMPASS shows superior performance within the context of surveillance for prescription drugs should provide strong encouragement of the promise of the active drug safety surveillance system.

In "Integrating active drug safety surveillance analyses across a network of observational healthcare databases", the consistency of COMPASS estimates was evaluated across five disparate data sources. This study replicated the design as the first aim, measuring the operating characteristics of COMPASS when examining labeled events and negative control outcomes associated with ACE inhibitor exposure. COMPASS was applied across five databases to assess the performance of each source independently, as well as alternative strategies for composite assessments across the data network. COMPASS was observed to have high specificity across all databases (>=0.97) and consistent positive predictive value (>=0.30), but with substantial differences in sensitivity (range: 0.04 to 0.45). Several differences among the data sources, such as population size; patient demographics and underlying disease severity; and longitudinality of data capture, may explain some of the performance inconsistency. The variability in performance characteristics across data sources should provide caution to those looking to generalize methodological results to a wide array of data sources. This study finding suggests each contributing source within a data network should be properly benchmarked through some retrospective empirical evaluation so as to gain sufficient understanding in how method results should be interpreted in the context of other findings.

Perhaps the most noteworthy observation from the study was the magnitude of heterogeneity that existed across sources when evaluating specific outcomes and its apparent impact on alternative strategies to synthesizing evidence across a data network. 82% of the statistically significant outcomes were observed to have high heterogeneity ($I^2 > 75\%$) of point estimates among databases. As a result, composite summaries based on both fixed-effects and random-effects meta-analysis of source-specific effect estimates did not yield additional predictive ability or identify additional outcomes not found by individual sources alone. The results suggest that, in the face of substantial heterogeneity, review should focus on the source-specific estimates and the explanation for why sources demonstrate consistency. An alternative approach to assess outcomes is on the basis of how many sources yield statistically significant results, following the principle that repeated independent replication should provide higher confidence in results. This study demonstrated that positive predictive value and specificity could be improved through increasingly restrictive criteria requiring 2+, 3+ or 4+ significant findings from the five databases. However, the number of pairs that satisfied these criteria decreased as well, reducing the sensitivity. Further independent replication was not sufficient to eliminate the risk of false positive findings, potentially due to consistent sources of bias that persisted throughout the analyses. These findings suggest that while highly consistent results across the network may be more reliable, it appears likely that a more common occurrence will be inconsistent estimates that are more difficult to discriminate.

These findings should give pause to the current trajectory of development of the national active surveillance system. A unique opportunity within a national system is the ability to examine the effects of medical products from across a network of disparate data sources, with the presumption that the source-specific estimates could be somehow combined to provide a more comprehensive summary. By pooling populations across the network to comprise over 100 million persons, it has been expected that sample size increases for exposed patients would enable more precise estimates of effects and facilitate exploration of rare events that are challenging to study in one source alone. The Mini-Sentinel protocol that evaluates the cardiovascular effects of saxagliptin and other oral anti-diabetic treatments assumes summary counts from each participating site across the Mini-Sentinel data network will be aggregated at the central coordinating center before conducting a composite Poisson regression at defined time intervals, and is powered accordingly for this type of pooled analysis[250]. If the results from the present study were generalizable to the Mini-Sentinel protocol, there could be substantial concerns that the composite estimates produced could be biased and less accurate than review of source-specific estimates. Further research is needed to determine the most effective strategies for synthesizing evidence across disparate observational data sources, as traditional meta-analytic approaches based on inverse variance weighting are likely to be insufficient to meet the challenges of bias presented in these data.

In “Comparative safety of ACE inhibitors: Evaluating an active surveillance framework,” the differential effects across ingredients within the ACE inhibitor class were explored using COMPASS against the Thomson MarketScan Commercial Claims and Encounters database. This study offered the opportunity to apply the active surveillance method to seven products (lisinopril, benazepril, enalapril, ramipril, quinapril, captopril, and moexipril) to assess whether product labeling differences in adverse event reporting were true clinical phenomenon observable in an administrative claims database. We observed substantial variation in the populations exposed and patterns of use amongst the seven products. The product differences in FDA-approved indications and listed contraindications resulted in COMPASS’s automated heuristics selecting unique comparator cohorts for each product and demonstrating differential success in covariate balance through propensity score stratification. Most risks were comparable across the ACE inhibitor class, with differences in product labeling not observed in real-world study. Two hypotheses were generated that suggest the risk of hypotensive outcomes of ramipril and enalapril may be elevated, though further exploration would be necessary to determine if this is a true causal effect. Also of note was that adverse events listed on the ACE inhibitor product labels did not appear to occur more frequently during ACE inhibitor exposure than the comparator groups.

This comparative safety assessment highlights the current gap in available evidence about relative safety effects of medical products. While product labels offer evidence of the occurrence of adverse events during the product lifecycle, they do not communicate the level of confidence in a causal attribution of the effect nor are they intended to communicate comparative differences between alternative treatments for the same indication. National efforts to develop active drug safety surveillance and comparative effectiveness systems have a significant opportunity to enable the establishment of a centralized source of real-world evidence about relative effects of alternative therapies, so that patients and providers can have a greater understanding of the potential outcomes in treatment.

### 7.3. Lessons through the evolution of the research program

The three manuscripts represent a summary of the findings that were generated as a final work product from the research. However, the body of work reflects an evolution in development from its inception. The proposed research design was followed to specification where possible, but necessary adjustments that came from the exploratory process are worth mentioning.

The most substantial area of improvement came in the iterative development of the COMPASS algorithm itself. Several enhancements were incorporated into COMPASS to address apparent limitations impacting the method's performance. In an attempt to increase the balance between cohorts on important potential confounders, the set of covariates used in the propensity score model was expanded to include lifestyle risk factors (obesity, tobacco, alcohol and drug abuse) and comorbid diseases included within the Charlson index. The lifestyle covariates were initially excluded due to the known limitations of administrative claims data in observing these effects. However, even if available data elements were poor proxies for these risk factors, it could be still be valuable to provide some level of adjustment in light of the strong confounding these variables can often induce with specific outcomes. The inclusion of the individual diseases within the Charlson index ensured greater balance among specific comorbidities in addition to balance of the composite index. COMPASS was also enhanced to allow both restriction and adjustment of indications and contraindications, with the recognition that either or both forms of control may be considered when exploring a given medical product. The current study's application of restriction on both indication and contraindication came at the recommendation of the FDA, who are most focused on managing risks for medical products when used as recommended, as opposed to unintended effects during off-label use.

A key lesson in the application of COMPASS that is not discussed in the study findings is the sensitivity of time-at-risk definition on effect estimates and method performance. As part of the proposed heuristic, COMPASS generates an effect estimate by selecting the maximum risk observed across multiple alternative risk windows (acute, subacute, insidious, and delayed). This was proposed under the premise that different adverse events have varied time-to-event relationships with exposure, which may not be fully characterized at the time of initial exploration. All empirical studies showed that COMPASS performance was optimized by focusing instead on the acute risk window, defined as 30 days from exposure initiation. This finding warrants further examination to determine if this is a consistent phenomenon or an artifact of this particular study design. It could be hypothesized that the acute risk window yielded the highest accuracy simply because the test cases under study (labeled events for ACE inhibitors) were more likely to have been acute time-to-event relationships. Alternatively, it is possible that using longitudinal healthcare data is more accurate during the immediate periods around exposure, and studies of effects requiring a longer duration between exposure and outcome are inherently more challenging due to the increasingly potential for confounding and noise in the data to surface.

The second study that examined COMPASS across a network of data sources was originally intended to include results from COMPASS against the Regenstrief Indiana Network of Patient Care (INPC) database, which would have provided an additional clinical source to complement the findings from the GE Centricity source. COMPASS was developed within the OMOP research lab, which afforded the opportunity to have access to de-identified patient-level data for the five data sources used in the study. The highly iterative process of methods development and evaluation demanded access to these data to support the exploration. Results from INPC were not included because the method execution within a distributed network framework presented additional logistical challenges that made the same level of exploration and understanding inefficient and unobtainable. The data access model currently advocated for active drug safety surveillance is a distributed network, whereby data holders maintain secure access to patient-level data and a central coordinating center is responsible for managing participation and aggregating summary-level results from across the network[140, 143, 244]. While this model has the potential to offer the high level of patient data privacy and foster more active participation among organizations who see their data holdings as proprietary, it presents a legitimate obstacle to methods development and evaluation. In-depth understanding of method performance often requires exploration of patient-level data to examine potential sources of unadjusted confounding or other previously unidentified artifacts in the data that can be biasing results. Because of the observed heterogeneity across disparate sources, it is likely unsafe to assume a method implementation on one data source is sufficiently generalizable to address the particularities of another source. The centralized data access model, where disparate sources are de-identified and made accessible through a common systems infrastructure, is more conducive to transparency and enabling full data exploration but raises its own set of concerns with data sharing. Until advanced analytics within an automated process can be demonstrated to be executed across a distributed network and yield reliable, accurate results, it seems reasonable to consider the choice of a data access model to be another outstanding question that requires further research.

The original study design called for examination of all ingredients within the ACE inhibitor class for the third aim. Beyond the seven products that were studied (lisinopril, benazepril, enalapril, ramipril, quinapril, captopril, and moexipril), this would have included fosinopril and perindopril. Perindopril was the least commonly prescribed ACE inhibitor across all data sources, while the number of fosinopril exposures was comparable to that of captopril. However, the automated heuristics in COMPASS were unable to be applied to these two products because the information source used for comparator selection did not include these ingredients in its set of relations. This limitation underscores the need for constant manual review throughout the automated systematic process envisioned for an active surveillance system. While the COMPASS algorithm was successfully applied in some circumstances, it is not feasible in other contexts, so it is important to determine the scenarios where the system will be unavailable to identify areas that require further supplemental effort.

### 7.4. Limitations of the COMPASS method

While the study results demonstrate promise for COMPASS as a viable active surveillance method, several limitations in the approach bear consideration for future enhancement. COMPASS applies an automated heuristic to select alternative treatments that serve as a comparator group. This comparator group is intended as a proxy for standard of care, in that it is constructed based on the set of medical products that could have been used for the target drug indication. This comparator group is used as a benchmark for assessing the relative effects of all potential outcomes. The heuristic, selecting all drugs that share an FDA-approved indication but have a different mechanism of action, is an objective approximation of the expert-based selection typically required in a customized evaluation study. Since the comparator can be comprised of multiple drugs, it is possible there is exists heterogeneity in effects amongst the comparator drugs that could results in the background rate being divergent from the true effects within any given comparator product. This may be particularly true for medical products with multiple indications, in which case it may be reasonable to consider stratifying the analysis by each indicated condition. It is also possible that the comparator may be inappropriate for certain outcomes, based on secondary indications or other factors. Further research is needed to assess the concordance of comparator selection between what would have been chosen by experts as compared to automated heuristics, and to assess how different choices would impact accuracy of method performance.

COMPASS uses a standardized procedure for covariate adjustment through propensity score stratification. As with all applications in propensity score adjustment, it is important to assess balance in baseline characteristics. In customized studies, when balance is insufficiently achieved, analysts may modify the propensity model and re-assess the adjustment approach. Within an automated process such as COMPASS, it is important to provide a comprehensive summary of covariate balance. While attempts have been made in this work and illustrated within the studies, it is possible that there is residual confounding due to insufficient control which would impact the interpretation of the effect estimates. While the COMPASS model uses proxies for demographics, lifestyle risk factors, comorbidities and health service utilization, it is possible that other covariates are relevant which have not been included in the model. For example, there may be additional covariates to consider that reflect diseases which are highly related to the indicated conditions. Some covariates used may be poorly recorded in specific databases, such as obesity and tobacco use in administrative claims. There are additional factors that may influence treatment selection that COMPASS doesn't account for because they are unavailable in the data model, such as patient socioeconomic status and provider-level characteristics.

COMPASS was implemented to be an efficient tool to facilitate rapid monitoring of medical products within an active surveillance network. Within a cohort design, it becomes very straightforward to simultaneously assess multiple outcomes for a given treatment. Specific attention was made to develop COMPASS to efficiently explore multiple treatments concurrently, but the nature of the design makes this operation more computationally demanding. One substantial value of the automated process is that multiple alternative design decisions can be evaluated simultaneously as part of a comprehensive sensitivity analysis. However, further work is needed to determine how to interpret results from across the sensitivity analysis, particularly when inconsistent findings are observed among seemingly reasonable parameter settings.

### 7.5. Limitations of observational data

A key limitation of this study is the fundamental challenge facing the enterprises of active drug safety surveillance and comparative effectiveness-- the integrity of the observational healthcare databases. In these studies, we evaluated an analytic method in its ability to identify temporal associations between drug exposure and outcome occurrence. However, the secondary use of administrative claims and electronic health record data requires an array of assumptions to infer drug exposure and the temporal relationship to disease onset. In neither type of data is there direct information about exposure, but instead information about prescriptions written by providers or dispensed by pharmacies. Outcomes are inferred from diagnosis codes, either captured from billing claims as part of reimbursement justification or from problem lists recorded by clinicians to support their care process. Typical pharmacoepidemiology evaluation studies may define outcomes using combinations of diagnosis codes, potentially in conjunction with other markers such as procedure codes or laboratory values, and often perform some level of source record verification to increase the confidence that observed events are true outcomes. In this study, individual diagnosis codes were used as crude proxies for outcome occurrence, with no source record verification. These diagnosis codes also serve to define the covariates used in the propensity adjustment and as the restriction criteria for indications and contraindications. COMPASS allows for cohort restriction based on prior indication recorded, but patients may have the indicated disease without having the diagnosis code recorded on a claim or have the indication show up in the record after exposure. This limitation of the underlying data could be a central reason for the overall poor performance of all surveillance methods and may provide one of the most significant opportunities for performance enhancements independent from the statistical methodology. As more robust electronic medical records are established and disparate data sources (claims, EHRs, personal health records, registries, death indexes, clinical trials) are able to be linked through common patient identifiers, not only will more comprehensive data be captured for individuals but also analyses should be able to more accurately ascertain patterns across populations.

### 7.6. Limitations of the COMPASS experiments

The studies conducted have provided initial evidence of the utility of COMPASS as a viable active surveillance method, but further studies should be designed to address some of the limitations in this existing work. Perhaps most impactful is that in order to justify the use of COMPASS as a reliable tool, it is important to have confidence that prior methodological research is generalizable to the types of scenarios anticipated by the envisioned active surveillance system. The current studies have focused on the performance of COMPASS within one class of medical products, ACE inhibitors, but it remains to be seen whether the operating characteristics observed are consistent with expectations for other prescription drugs. Focus on one drug class has allowed for a deeper dive and firmer understanding of how the method behaved, but limits the overall generalizability. OMOP has created a larger panel of drug-outcome pairs to study but suffers from problem of breadth vs. depth. The knowledge needed to have confidence in the creation of a national system requires both breadth and depth in methodological research. This study is only a first step in that direction. Evaluation of method performance in its accuracy to discriminate between positive controls and negative controls rests on the confidence that the ground truth established for the test cases is in fact accurate. In this study, positive controls were defined by adverse events that were listed in the product labeling for ACE inhibitors, while negative controls were selected based on conditions that were unrelated to any known drug effect. Given the maturity and wide use of the drug class, it seems reasonable to expect the negative controls to be accurate, with little chance that a true effect exists among them that had not been previously discovered. However, as the third study showed, some adverse events listed on product labels are not necessarily true causal effects and therefore shouldn't expect to be observed as positive associations. As a result, method performance may be understating sensitivity if some of the test cases are misclassified as true effects. Future experiments should establish a reference set where all test cases have high confidence in correct classification of causal status to minimize this concern of measurement error. Use of simulated data, where ground truth can be defined a priori, may be a valuable supplement to these real-world investigations.

COMPASS was evaluated on its overall performance as a single tool for active surveillance. However, it is quite possible that stakeholders should not expect a single method to be a magic bullet with consistently reliable performance, but instead should consider that the tool's accuracy may vary by attributes of the drug and outcome under study or based on the database that it is applied against. It could be that some methods are more or less appropriate for specific circumstances, but determining these scenario operating characteristics cannot be judged through expert subjective assessment alone and requires further empirical research.

COMPASS was evaluated against five disparate databases, which represents one of the most comprehensive methodological assessments for drug safety to date. However, given the substantial heterogeneity that was observed across those five sources, the study raises questions about the generalizability of the findings to other data sources. It could be reasonable to expect that performance results would be different had COMPASS been applied to a different network of databases. Further expansion of this methodological research to additional data sources should begin to yield insights about the sources of heterogeneity and how to predict method accuracy across a broader range of available data resources.

### 7.7. Contributions to the field

With these limitations in mind, these studies provide a solid foundation of research that should directly inform an important national issue.  The results of this study should inform decisions about the appropriateness and utility of analyzing observational data as part of a future drug safety surveillance process and add to the literature in several important ways, with clinical, methodological, and policy implications.

First, from the clinical perspective, the exploratory analyses of ACE inhibitors have provided the first known comprehensive assessment of the comparative safety of ACE inhibitors in an active surveillance framework.  In light of the comparable safety profiles, there may be interest in examining why these products have inconsistencies in product labeling and how further comparative studies can better inform clinical practice about appropriate use of products within the class.

Second, from a methodological perspective, the study has detailed and provided empirical evidence to inform the potential use of a novel method for identifying drug safety issues in automated healthcare databases as part of an active surveillance system.  This method leverages advances in pharmacoepidemiology, biomedical informatics, and pharmaceutical sciences to provide an analytical framework that could support continued drug outcome research beyond the scope of this study's ACE inhibitor analyses.

Finally, from a policy perspective, the evaluation of how to interpret findings across a network of data sources may have broader implications for initiating the national active surveillance system. There is little research to inform how decision-making processes will accommodate information when generating, strengthening and confirming hypotheses about potential drug-related effects[23]. Prior to this study, the role of exploratory analyses in an active surveillance system and the relative confidence in information that can be gained from such analyses was undetermined. The examination of heterogeneity across sources and the potential use of a meta-analytic framework to integrate estimates have provided insights that should inform the governance of the future national active surveillance system. More broadly, the measurement of operating characteristics of an active surveillance system should help establish a greater understanding of how to interpret surveillance results in the context of all other available information as part of a causality assessment for an emerging drug safety issue.

While this body of work represents a significant contribution to the field, it is a small step on a long-term journey toward developing a capability for improving our understanding of the effects of medical products. This research has raised many additional questions that warrant further investigation. Improving the performance of methods requires a deep-dive exploration to better explain why methods failed to identify known effects or falsely highlighted positive associations for negative controls, so that strategies can be developed to mitigate these inaccuracies. This exploration requires further examination of other potential sources of bias and testing hypotheses about how inclusion of additional covariates and/or imposing new inclusion/exclusion criteria impacts effect estimates. Such work demands clinical review of patient-level records and secondary confirmation that operational definitions for exposure and outcome are providing appropriate ascertainment across the data sources. Methods improvement needs to be complemented with more precise and comprehensive evaluations of performance through the expansion of methodological experiments to include a broader set of medical products and outcomes that better reflect the anticipated scenarios envisioned within a comparative effectiveness and active surveillance system. A sustainable research partnership that supports a common experimental framework is paramount to establishing a benchmark for current expectations, facilitating discovery and development of methodological innovation, and measuring progress as research continues moving forward.

Developing a high-quality system for evidence development using a network of observational healthcare databases requires active participation from all stakeholders, including government, industry, academia, and health care organizations, and it demands a full understanding of the perspectives from the decision-makers that these analyses should ultimately inform, including regulators, payers, providers, and patients. Advancing the science of active surveillance and clinical effectiveness requires interdisciplinary collaboration between statistics, epidemiology, health services research, computer science, medical informatics, engineering, and the clinical sciences. As these innovations are developed and applied in medical practice, special attention will continue to be needed to ensure the appropriate use of electronic healthcare data and interpretation of inferences about the effects of medical treatments.

# WORKS CITED

**1.** Berlin JA, Glasser SC, Ellenberg SS. Adverse event detection in drug development: recommendations and obligations beyond phase 3. *Am J Public Health.* Aug 2008;98(8):1366-1371.

**2.** Knapp P, Raynor DK, Berry DC. Comparison of two methods of presenting risk information to patients about the side effects of medicines. *Qual Saf Health Care.* Jun 2004;13(3):176-180.

**3.** Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA.* Apr 15 1998;279(15):1200-1205.

**4.** Bennett CL, Nebeker JR, Yarnold PR, et al. Evaluation of serious adverse drug reactions: a proactive pharmacovigilance program (RADAR) vs safety activities conducted by the Food and Drug Administration and pharmaceutical manufacturers. *Arch Intern Med.* May 28 2007;167(10):1041-1049.

**5.** Furberg CD, Levin AA, Gross PA, Shapiro RS, Strom BL. The FDA and drug safety: a proposal for sweeping changes. *Arch Intern Med.* Oct 9 2006;166(18):1938-1942.

**6.** *Managing the Risks From Medical Product Use: Creating a Risk Management Framework.* : US Department of Health and Human Services, Food and Drug Administration; May 1999 1999.

**7.** Lasser KE, Allen PD, Woolhandler SJ, Himmelstein DU, Wolfe SM, Bor DH. Timing of new black box warnings and withdrawals for prescription medications. *JAMA.* May 1 2002;287(17):2215-2220.

**8.** Almenoff J, Tonning JM, Gould AL, et al. Perspectives on the use of data mining in pharmacovigilance. *Drug Saf.* 2005;28(11):981-1007.

**9.** DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *American Statistician.* 1999;53(3):177-189.

**10.** Szarfman A, Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. *Drug Saf.* 2002;25(6):381-392.

**11.** Public Law 110-85: Food and Drug Administration Amendments Act of 2007; 2007.

**12.** Racoosin J. FDA's Sentinel Initiative — A National Strategy for Monitoring Medical Product Safety. *2nd Drug Information Association (DIA) Conference on Signal Detection and Data Mining*. New York, NY; 2009.

**13.** Schneeweiss S. On Guidelines for Comparative Effectiveness Research Using Nonrandomized Studies in Secondary Data Sources. *Value Health.* Sep 10 2009.

**14.** Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol.* Apr 2005;58(4):323-337.

**15.** Swets JA. Measuring the accuracy of diagnostic systems. *Science.* Jun 3 1988;240(4857):1285-1293.

**16.** Ryan PB. Review of Observational Analysis Methods. http://omop.fnih.org/?q=node/61. Accessed April 6, 2009.

**17.** Nelson J, Cook A, Yu O. Evaluation of signal detection methods for use in prospective postlicensure medical product safety surveillance. http://www.fda.gov/OHRMS/DOCKETS/98fr/FDA-2009-N-0192-rpt.pdf. Accessed 8 May 2009.

**18.** Ryan PB, Powell, G.E., Pattishall, E.N., Beach, K.J. Performance of Screening Multiple Observational Databases for Active Drug Safety Surveillance. *International Society of Pharmacoepidemiology*. Providence, RI; 2009.

**19.** Whitaker HJ, Hocine MN, Farrington CP. The methodology of self-controlled case series studies. *Stat Methods Med Res.* Feb 2009;18(1):7-26.

**20.** Whitaker HJ, Farrington CP, Spiessens B, Musonda P. Tutorial in biostatistics: the self-controlled case series method. *Stat Med.* May 30 2006;25(10):1768-1797.

**21.** Morton S, Adams J, Suttorp M, Shekelle P. *Meta-regression Approaches: What, Why, When, and How? Technical Review 8 (Prepared by Southern California–RAND Evidence-based Practice Center, under Contract No 290-97-0001). AHRQ Publication No. 04-0033.* Rockville, MD: Agency for Healthcare Research and Quality; March 2004 2004.

**22.** Chou R, Helfand M, Carson S. *Drug Class Review on Angiotensin Converting Enzyme Inhibitors* 1995.

**23.** Avorn J, Schneeweiss S. Managing drug-risk information--what to do with all those new numbers. *N Engl J Med.* Aug 13 2009;361(7):647-649.

**24.** Burton MM, Hope C, Murray MD, Hui S, Overhage JM. The cost of adverse drug events in ambulatory care. *AMIA Annu Symp Proc.* 2007:90-93.

**25.** Fontanarosa PB, Rennie D, DeAngelis CD. Postmarketing surveillance--lack of vigilance, lack of trust. *JAMA.* Dec 1 2004;292(21):2647-2650.

**26.** Avorn J. Evaluating drug effects in the post-Vioxx world: there must be a better way. *Circulation.* May 9 2006;113(18):2173-2176.

**27.** FDA. VIOXX® (rofecoxib tablets and oral suspension). http://www.fda.gov/cder/foi/label/2004/021052s026_021042s019lbl.pdf. Accessed April 26, 2009, 2009.

**28.** FDA. Merck Withdraws Vioxx; FDA Issues Public Health Advisory. http://www.fda.gov/fdac/features/2004/604_vioxx.html. Accessed April 26, 2009.

**29.** FDA. FDA Public Health Advisory: Safety of Vioxx. http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm106274.htm. Accessed July 8, 2009.

**30.** FDA. FDA Public Health Advisory: Tegaserod maleate (marketed as Zelnorm) http://www.fda.gov/Drugs/DrugSafety/PublicHealthAdvisories/ucm051284.html. Accessed July 8, 2009.

**31.** FDA. Alert for Healthcare Professionals: Pemoline Tablets and Chewable Tablets (marketed as Cylert). http://www.fda.gov/Cder/drug/InfoSheets/HCP/pemolineHCP.htm. Accessed April 29, 2009.

**32.** FDA. Labeling Revised for Diabetes Drug Avandia. http://www.fda.gov/ForConsumers/ConsumerUpdates/ucm049058.htm. Accessed July 9, 2009.

**33.** Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *N Engl J Med.* Jun 14 2007;356(24):2457-2471.

**34.** Psaty BM, Furberg CD. The record on rosiglitazone and the risk of myocardial infarction. *N Engl J Med.* Jul 5 2007;357(1):67-69.

**35.** Kazi D. Rosiglitazone and implications for pharmacovigilance. *BMJ.* Jun 16 2007;334(7606):1233-1234.

**36.** Routledge P. 150 years of pharmacovigilance. *Lancet.* Apr 18 1998;351(9110):1200-1201.

**37.** van Grootheest K. The Dawn of Pharmacovigilance: An Historical Perspective. *Int J Pharm Med.* 2003(17 (5-6)):195-200.

**38.** Kessler DA. Introducing MEDWatch. A new approach to reporting medication and device adverse effects and product problems. *JAMA.* Jun 2 1993;269(21):2765-2768.

**39.** Meyboom RH, Egberts AC, Gribnau FW, Hekster YA. Pharmacovigilance in perspective. *Drug Saf.* Dec 1999;21(6):429-447.

**40.** Gough S. Post-marketing surveillance: a UK/European perspective. *Curr Med Res Opin.* Apr 2005;21(4):565-570.

**41.** Waller PC, Evans SJ. A model for the future conduct of pharmacovigilance. *Pharmacoepidemiol Drug Saf.* Jan-Feb 2003;12(1):17-29.

**42.** Baciu A, Stratton K, Burke S. *The Future of Drug Safety: Promoting and Protecting the Health of the Public*: Institute of Medicine; 2006.

**43.** Hennessy S, Strom BL. PDUFA reauthorization--drug safety's golden moment of opportunity? *N Engl J Med.* Apr 26 2007;356(17):1703-1704.

**44.** FDA. The Sentinel Initiative: A National Strategy for Monitoring Medical Product Safety. May 2008; http://www.fda.gov/Safety/FDAsSentinelInitiative/ucm089474.htm.

**45.** Rosenbaum P. *Observational Studies*: Springer; 2002.

**46.** Egger M, Smith GD, Altman D. *Systematic Reviews in Health Care: Meta-analysis in context*: BMJ Publishing Group; 2001.

**47.** Higgins J, Green S. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2*: The Cochrane Collaboration; 2009.

**48.** Ross JS, Madigan D, Hill KP, Egilman DS, Wang Y, Krumholz HM. Pooled analysis of rofecoxib placebo-controlled clinical trial data: lessons for postmarket pharmaceutical safety surveillance. *Arch Intern Med.* Nov 23 2009;169(21):1976-1985.

**49.** Shrier I, Boivin JF, Platt RW, et al. The interpretation of systematic reviews with meta-analyses: an objective or subjective process? *BMC Med Inform Decis Mak.* 2008;8:19.

**50.** Foody JM, Mendys PM, Liu LZ, Simpson RJ, Jr. The utility of observational studies in clinical decision making: lessons learned from statin trials. *Postgrad Med.* May;122(3):222-229.

**51.** Hartzema AG, Porta MS, Tilson HH. *Pharmacoepidemiology: An Introduction*. 3 ed. Cincinnati, OH: Harvey Whitney Books; 1999.

**52.** Hartzema AG, Tilson HH, Chan KA. *Pharmacoepidemiology and Therapeutic Risk Management*. Cincinnati, OH: Harvey Whitney Books; 2008.

**53.** Rothman K. *Epidemiology: An Introduction*: Oxford University Press; 2002.

**54.** Rothman K, Greenland S, Lash T. *Modern Epidemiology*: Lippincott Williams & Wilkins; 2008.

**55.** Szklo M, Nieto FJ. *Epidemiology: Beyond the Basics*: Jones and Bartlett Publishers; 2007.

**56.** Jewell N. *Statistics for epidemiology*: Chapman & Hall; 2004.

**57.** Strom B. *Pharmacoepidemiology*. 4 ed. Chichester, UK: Wiley; 2005.

**58.** Rodriguez EM, Staffa JA, Graham DJ. The role of databases in drug postmarketing surveillance. *Pharmacoepidemiol Drug Saf.* Aug-Sep 2001;10(5):407-410.

**59.** Rockhill B, Newman B, Weinberg C. Use and misuse of population attributable fractions. *Am J Public Health.* Jan 1998;88(1):15-19.

**60.** Weatherby LB, Nordstrom BL, Fife D, Walker AM. The impact of wording in "Dear doctor" letters and in black box labels. *Clin Pharmacol Ther.* Dec 2002;72(6):735-742.

**61.** Hennessy S. Use of health care databases in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol.* Mar 2006;98(3):311-313.

**62.** Suissa S, Garbe E. Primer: administrative health databases in observational studies of drug effects--advantages and disadvantages. *Nat Clin Pract Rheumatol.* Dec 2007;3(12):725-732.

**63.** Polinski JM, Schneeweiss S, Levin R, Shrank WH. Completeness of retail pharmacy claims data: implications for pharmacoepidemiologic studies and pharmacy practice in elderly patients. *Clin Ther.* Sep 2009;31(9):2048-2059.

**64.** Short PF, Graefe DR, Schoen C. *Churn, Churn, Churn: How Instability of Health Insurance Shapes America's Uninsured Problem*: The Commonwealth Fund; 2003.

**65.** Lewis JD, Brensinger C. Agreement between GPRD smoking data: a survey of general practitioners and a population-based survey. *Pharmacoepidemiol Drug Saf.* Jul 2004;13(7):437-441.

**66.** Hennessy S, Leonard CE, Freeman CP, et al. Validation of diagnostic codes for outpatient-originating sudden cardiac death and ventricular arrhythmia in Medicaid and Medicare claims data. *Pharmacoepidemiol Drug Saf.* Oct 20 2009.

**67.** Wahl PM, Rodgers K, Schneeweiss S, et al. Validation of claims-based diagnostic and procedure codes for cardiovascular and gastrointestinal serious adverse events in a commercially-insured population. *Pharmacoepidemiol Drug Saf.* Feb 5.

**68.** Donahue JG, Weiss ST, Goetsch MA, Livingston JM, Greineder DK, Platt R. Assessment of asthma using automated and full-text medical records. *J Asthma.* 1997;34(4):273-281.

**69.** Lee DS, Donovan L, Austin PC, et al. Comparison of coding of heart failure and comorbidities in administrative and clinical data for use in outcomes research. *Med Care.* Feb 2005;43(2):182-188.

**70.** Miller DR, Oliveria SA, Berlowitz DR, Fincke BG, Stang P, Lillienfeld DE. Angioedema incidence in US veterans initiating angiotensin-converting enzyme inhibitors. *Hypertension.* Jun 2008;51(6):1624-1630.

**71.** Pladevall M, Goff DC, Nichaman MZ, et al. An assessment of the validity of ICD Code 410 to identify hospital admissions for myocardial infarction: The Corpus Christi Heart Project. *Int J Epidemiol.* Oct 1996;25(5):948-952.

**72.** So L, Evans D, Quan H. ICD-10 coding algorithms for defining comorbidities of acute myocardial infarction. *BMC Health Serv Res.* 2006;6:161.

**73.** Tunstall-Pedoe H. Validity of ICD code 410 to identify hospital admission for myocardial infarction. *Int J Epidemiol.* Apr 1997;26(2):461-462.

**74.** Varas-Lorenzo C, Castellsague J, Stang MR, Tomas L, Aguado J, Perez-Gutthann S. Positive predictive value of ICD-9 codes 410 and 411 in the identification of cases of acute coronary syndromes in the Saskatchewan Hospital automated database. *Pharmacoepidemiol Drug Saf.* Aug 2008;17(8):842-852.

**75.** Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. *J Clin Epidemiol.* Feb 2004;57(2):131-141.

**76.** Garcia Rodriguez LA, Perez Gutthann S. Use of the UK General Practice Research Database for pharmacoepidemiology. *Br J Clin Pharmacol.* May 1998;45(5):419-425.

**77.** Leonard CE, Haynes K, Localio AR, et al. Diagnostic E-codes for commonly used, narrow therapeutic index medications poorly predict adverse drug events. *J Clin Epidemiol.* Jun 2008;61(6):561-571.

**78.** Harrold LR, Saag KG, Yood RA, et al. Validity of gout diagnoses in administrative data. *Arthritis Rheum.* Feb 15 2007;57(1):103-108.

**79.** Lewis JD, Schinnar R, Bilker WB, Wang X, Strom BL. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. *Pharmacoepidemiol Drug Saf.* Apr 2007;16(4):393-401.

**80.** Strom BL. Data validity issues in using claims data. *Pharmacoepidemiol Drug Saf.* Aug-Sep 2001;10(5):389-392.

**81.** Hennessy S, Leonard CE, Palumbo CM, Newcomb C, Bilker WB. Quality of Medicaid and Medicare data obtained through Centers for Medicare and Medicaid Services (CMS). *Med Care.* Dec 2007;45(12):1216-1220.

**82.** Walker AM. Confounding by indication. *Epidemiology.* Jul 1996;7(4):335-336.

**83.** Bosco JL, Silliman RA, Thwin SS, et al. A most stubborn bias: no adjustment method fully resolves confounding by indication in observational studies. *J Clin Epidemiol.* Jan;63(1):64-74.

**84.** Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. *Med Care.* Jun;48(6 Suppl):S114-120.

**85.** Rothman KJ, Suissa S. Exclusion of immortal person-time. *Pharmacoepidemiol Drug Saf.* Oct 2008;17(10):1036.

**86.** Suissa S. Immortal time bias in pharmaco-epidemiology. *Am J Epidemiol.* Feb 15 2008;167(4):492-499.

**87.** Suissa S. Immortal time bias in observational studies of drug effects. *Pharmacoepidemiol Drug Saf.* Mar 2007;16(3):241-249.

**88.** Schneeweiss S, Patrick AR, Sturmer T, et al. Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. *Med Care.* Oct 2007;45(10 Supl 2):S131-142.

**89.** Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol.* Nov 1 2003;158(9):915-920.

**90.** Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. *Pharmacoepidemiol Drug Saf.* 2010.

**91.** Cadarette SM, Katz JN, Brookhart MA, et al. Comparative gastrointestinal safety of weekly oral bisphosphonates. *Osteoporos Int.* Oct 2009;20(10):1735-1747.

**92.** Bravo G, Dubois MF, Hebert R, De Wals P, Messier L. A prospective evaluation of the Charlson Comorbidity Index for use in long-term care patients. *J Am Geriatr Soc.* Apr 2002;50(4):740-745.

**93.** Charlson M, Szatrowski TP, Peterson J, Gold J. Validation of a combined comorbidity index. *J Clin Epidemiol.* Nov 1994;47(11):1245-1251.

**94.** Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis.* 1987;40(5):373-383.

**95.** Cleves MA, Sanchez N, Draheim M. Evaluation of two competing methods for calculating Charlson's comorbidity index when analyzing short-term mortality using administrative data. *J Clin Epidemiol.* Aug 1997;50(8):903-908.

**96.** D'Hoore W, Bouckaert A, Tilquin C. Practical considerations on the use of the Charlson comorbidity index with administrative data bases. *J Clin Epidemiol.* Dec 1996;49(12):1429-1433.

**97.** D'Hoore W, Sicotte C, Tilquin C. Risk adjustment in outcome assessment: the Charlson comorbidity index. *Methods Inf Med.* Nov 1993;32(5):382-387.

**98.** Li B, Evans D, Faris P, Dean S, Quan H. Risk adjustment performance of Charlson and Elixhauser comorbidities in ICD-9 and ICD-10 administrative databases. *BMC Health Serv Res.* 2008;8:12.

**99.** Needham DM, Scales DC, Laupacis A, Pronovost PJ. A systematic review of the Charlson comorbidity index using Canadian administrative databases: a perspective on risk adjustment in critical care research. *J Crit Care.* Mar 2005;20(1):12-19.

**100.** Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care.* Nov 2005;43(11):1130-1139.

**101.** Southern DA, Quan H, Ghali WA. Comparison of the Elixhauser and Charlson/Deyo methods of comorbidity measurement in administrative data. *Med Care.* Apr 2004;42(4):355-360.

**102.** Zhang JX, Iwashyna TJ, Christakis NA. The performance of different lookback periods and sources of information for Charlson comorbidity adjustment in Medicare claims. *Med Care.* Nov 1999;37(11):1128-1139.

**103.** Farley JF, Harley CR, Devine JW. A comparison of comorbidity measurements to predict healthcare expenditures. *Am J Manag Care.* Feb 2006;12(2):110-119.

**104.** Schneeweiss S, Seeger JD, Maclure M, Wang PS, Avorn J, Glynn RJ. Performance of comorbidity scores to control for confounding in epidemiologic studies using claims data. *Am J Epidemiol.* Nov 1 2001;154(9):854-864.

**105.** Lunt M, Solomon D, Rothman K, et al. Different methods of balancing covariates leading to different effect estimates in the presence of effect modification. *Am J Epidemiol.* Apr 1 2009;169(7):909-917.

**106.** Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med.* Oct 15 1997;127(8 Pt 2):757-763.

**107.** D'Agostino RB, Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med.* Oct 15 1998;17(19):2265-2281.

**108.** Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *Am J Epidemiol.* Jun 15 2006;163(12):1149-1156.

**109.** Sturmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol.* May 2006;59(5):437-447.

**110.** Glynn RJ, Schneeweiss S, Sturmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol.* Mar 2006;98(3):253-259.

**111.** Sturmer T, Glynn RJ, Rothman KJ, Avorn J, Schneeweiss S. Adjustments for unmeasured confounders in pharmacoepidemiologic database studies using external information. *Med Care.* Oct 2007;45(10 Supl 2):S158-165.

**112.** Schneeweiss S, Glynn RJ, Tsai EH, Avorn J, Solomon DH. Adjusting for unmeasured confounders in pharmacoepidemiologic claims data using external information: the example of COX2 inhibitors and myocardial infarction. *Epidemiology.* Jan 2005;16(1):17-24.

**113.** Seeger JD, Kurth T, Walker AM. Use of propensity score technique to account for exposure-related covariates: an example and lesson. *Med Care.* Oct 2007;45(10 Supl 2):S143-148.

**114.** Seeger JD, Walker AM, Williams PL, Saperia GM, Sacks FM. A propensity score-matched cohort study of the effect of statins, mainly fluvastatin, on the occurrence of acute myocardial infarction. *Am J Cardiol.* Dec 15 2003;92(12):1447-1451.

**115.** Seeger JD, Williams PL, Walker AM. An application of propensity score matching using claims data. *Pharmacoepidemiol Drug Saf.* Jul 2005;14(7):465-476.

**116.** Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology.* Jul 2009;20(4):512-522.

**117.** Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Saf.* May 2006;15(5):291-303.

**118.** Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf.* Mar 30.

**119.** Hogan JW, Lancaster T. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Stat Methods Med Res.* Feb 2004;13(1):17-48.

**120.** Schneeweiss S. Developments in post-marketing comparative effectiveness research. *Clin Pharmacol Ther.* Aug 2007;82(2):143-156.

**121.** Dudl RJ, Wang MC, Wong M, Bellows J. Preventing myocardial infarction and stroke with a simplified bundle of cardioprotective medications. *Am J Manag Care.* Oct 2009;15(10):e88-94.

**122.** Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. *J Clin Epidemiol.* Dec 2009;62(12):1226-1232.

**123.** Rassen JA, Mittleman MA, Glynn RJ, Alan Brookhart M, Schneeweiss S. Safety and effectiveness of bivalirudin in routine care of patients undergoing percutaneous coronary intervention. *Eur Heart J.* Mar;31(5):561-572.

**124.** Schneeweiss S, Seeger JD, Landon J, Walker AM. Aprotinin during coronary-artery bypass grafting and risk of death. *N Engl J Med.* Feb 21 2008;358(8):771-783.

**125.** Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology.* Jul 2009;20(4):488-495.

**126.** Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. *Int J Epidemiol.* Apr;39(2):417-420.

**127.** Platt R, Davis R, Finkelstein J, et al. Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics. *Pharmacoepidemiol Drug Saf.* Aug-Sep 2001;10(5):373-377.

**128.** Andrade SE, Graham DJ, Staffa JA, et al. Health plan administrative databases can efficiently identify serious myopathy and rhabdomyolysis. *J Clin Epidemiol.* Feb 2005;58(2):171-174.

**129.** Andrade SE, Majumdar SR, Chan KA, et al. Low frequency of treatment of osteoporosis among postmenopausal women following a fracture. *Arch Intern Med.* Sep 22 2003;163(17):2052-2057.

**130.** Andrade SE, Raebel MA, Brown J, et al. Outpatient use of cardiovascular drugs during pregnancy. *Pharmacoepidemiol Drug Saf.* Mar 2008;17(3):240-247.

**131.** Andrade SE, Raebel MA, Morse AN, et al. Use of prescription medications with a potential for fetal harm among pregnant women. *Pharmacoepidemiol Drug Saf.* Aug 2006;15(8):546-554.

**132.** Davis RL, Rubanowice D, McPhillips H, et al. Risks of congenital malformations and perinatal events among infants exposed to antidepressant medications during pregnancy. *Pharmacoepidemiol Drug Saf.* Oct 2007;16(10):1086-1094.

**133.** Harrold LR, Andrade SE, Eisner M, et al. Identification of patients with Churg-Strauss syndrome (CSS) using automated data. *Pharmacoepidemiol Drug Saf.* Oct 2004;13(10):661-667.

**134.** Harrold LR, Andrade SE, Go AS, et al. Incidence of Churg-Strauss syndrome in asthma drug users: a population-based perspective. *J Rheumatol.* Jun 2005;32(6):1076-1080.

**135.** Harrold LR, Patterson MK, Andrade SE, et al. Asthma drug use and the development of Churg-Strauss syndrome (CSS). *Pharmacoepidemiol Drug Saf.* Jun 2007;16(6):620-626.

**136.** Velentgas P, Bohn RL, Brown JS, et al. A distributed research network model for post-marketing safety studies: the Meningococcal Vaccine Study. *Pharmacoepidemiol Drug Saf.* Dec 2008;17(12):1226-1234.

**137.** Brown JS, Kulldorff M, Chan KA, et al. Early detection of adverse drug events within population-based health networks: application of sequential testing methods. *Pharmacoepidemiol Drug Saf.* Dec 2007;16(12):1275-1284.

**138.** Brown JS, Kulldorff M, Petronis KR, et al. Early adverse drug event signal detection within population-based health networks using sequential methods: key methodologic considerations. *Pharmacoepidemiol Drug Saf.* Mar 2009;18(3):226-234.

**139.** Rassen JA, Avorn J, Schneeweiss S. Multivariate-adjusted pharmacoepidemiologic analyses of confidential information pooled from multiple health care utilization databases. *Pharmacoepidemiol Drug Saf.* Feb 16.

**140.** Brown J, Lane K, Moore K, Platt R. Defining and Evaluating Possible Database Models to Implement the FDA Sentinel Initiative http://www.fda.gov/OHRMS/DOCKETS/98fr/FDA-2009-N-0192-0005.pdf Accessed 30 Aug 2009.

**141.** Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Med Care.* Jun;48(6 Suppl):S45-51.

**142.** Holmes JH, Brown J, Hennessy S, et al. Developing a distributed research network to conduct population-based studies and safety surveillance. *AMIA Annu Symp Proc.* 2008:973.

**143.** Maro JC, Platt R, Holmes JH, et al. Design of a National Distributed Health Data Network. *Ann Intern Med.* Jul 28 2009.

**144.** FDA. Guidance for industry: good pharmacovigilance practices and pharmacoepidemiologic assessment. In: Research UFaDACfDEaRaCfBEa, ed; 2005.

**145.** Hauben M, Patadia V, Gerrits C, Walsh L, Reich L. Data mining in pharmacovigilance: the need for a balanced perspective. *Drug Saf.* 2005;28(10):835-842.

**146.** Waller PC, Lee EH. Responding to drug safety issues. *Pharmacoepidemiol Drug Saf.* Dec 1999;8(7):535-552.

**147.** Psaty BM, Furberg CD. COX-2 inhibitors--lessons in drug safety. *N Engl J Med.* Mar 17 2005;352(11):1133-1135.

**148.** Seligman PJ. 'Dear doctor...'--evaluating the impact of risk communication efforts. *Pharmacoepidemiol Drug Saf.* Jun 2003;12(4):291-293.

**149.** Aronson JK, Hauben M. Anecdotes that provide definitive evidence. *BMJ.* Dec 16 2006;333(7581):1267-1269.

**150.** Horisberger B, Dinkel R. *The perception and management of drug safety risks*: Springer-Verlag; 1989.

**151.** Vandenbroucke JP. In defense of case reports and case series. *Ann Intern Med.* Feb 20 2001;134(4):330-334.

**152.** Noren GN, Edwards IR. Modern methods of pharmacovigilance: detecting adverse effects of drugs. *Clin Med.* Oct 2009;9(5):486-489.

**153.** Hauben M, Madigan D, Gerrits CM, Walsh L, Van Puijenbroek EP. The role of data mining in pharmacovigilance. *Expert Opin Drug Saf.* Sep 2005;4(5):929-948.

**154.** Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol Drug Saf.* Oct-Nov 2001;10(6):483-486.

**155.** Rothman KJ, Lanes S, Sacks ST. The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiol Drug Saf.* Aug 2004;13(8):519-523.

**156.** van Hunsel F, van Puijenbroek E, de Jong-van den Berg L, van Grootheest K. Media attention and the influence on the reporting odds ratio in disproportionality analysis: an example of patient reporting of statins. *Pharmacoepidemiol Drug Saf.* Jan;19(1):26-32.

**157.** van Puijenbroek E, Diemont W, van Grootheest K. Application of quantitative signal detection in the Dutch spontaneous reporting system for adverse drug reactions. *Drug Saf.* 2003;26(5):293-301.

**158.** Almenoff JS, Pattishall EN, Gibbs TG, DuMouchel W, Evans SJ, Yuen N. Novel statistical tools for monitoring the safety of marketed drugs. *Clin Pharmacol Ther.* Aug 2007;82(2):157-166.

**159.** van Manen RP, Fram D, DuMouchel W. Signal detection methodologies to support effective safety management. *Expert Opin Drug Saf.* Jul 2007;6(4):451-464.

**160.** Hansen RA, Gartlehner G, Powell GE, Sandler RS. Serious adverse events with infliximab: analysis of spontaneously reported adverse events. *Clin Gastroenterol Hepatol.* Jun 2007;5(6):729-735.

**161.** Bate A. Bayesian confidence propagation neural network. *Drug Saf.* 2007;30(7):623-625.

**162.** Bate A, Edwards IR. Data mining in spontaneous reports. *Basic Clin Pharmacol Toxicol.* Mar 2006;98(3):324-330.

**163.** Bate A, Evans SJ. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiol Drug Saf.* Apr 8 2009.

**164.** Bate A, Lindquist M, Edwards IR. The application of knowledge discovery in databases to post-marketing drug safety: example of the WHO database. *Fundam Clin Pharmacol.* Apr 2008;22(2):127-140.

**165.** Bate A, Lindquist M, Edwards IR, et al. A Bayesian neural network method for adverse drug reaction signal generation. *Eur J Clin Pharmacol.* Jun 1998;54(4):315-321.

**166.** Bate A, Lindquist M, Edwards IR, Orre R. A data mining approach for signal detection and analysis. *Drug Saf.* 2002;25(6):393-397.

**167.** Bate A, Lindquist M, Orre R, Edwards IR, Meyboom RH. Data-mining analyses of pharmacovigilance signals in relation to relevant comparison drugs. *Eur J Clin Pharmacol.* Oct 2002;58(7):483-490.

**168.** Hopstadius J, Noren GN, Bate A, Edwards IR. Impact of stratification on adverse drug reaction surveillance. *Drug Saf.* 2008;31(11):1035-1048.

**169.** Noren GN, Bate A, Orre R, Edwards IR. Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. *Stat Med.* Nov 15 2006;25(21):3740-3757.

**170.** Noren GN, Sundberg R, Bate A, Edwards IR. A statistical methodology for drug-drug interaction surveillance. *Stat Med.* Jul 20 2008;27(16):3057-3070.

**171.** Hochberg A, Hauben M. Time-to-Signal Comparison for Drug Safety Data-Mining Algorithms vs. Traditional Signaling Criteria. *Clin Pharmacol Ther.* Mar 25 2009.

**172.** van Puijenbroek EP, Bate A, Leufkens HG, Lindquist M, Orre R, Egberts AC. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiol Drug Saf.* Jan-Feb 2002;11(1):3-10.

**173.** Almenoff JS, LaCroix KK, Yuen NA, Fram D, DuMouchel W. Comparative performance of two quantitative safety signalling methods: implications for use in a pharmacovigilance department. *Drug Saf.* 2006;29(10):875-887.

**174.** Hochberg AM, Hauben M, Pearson RK, et al. An evaluation of three signal-detection algorithms using a highly inclusive reference event database. *Drug Saf.* 2009;32(6):509-525.

**175.** Chan KA, Hauben M. Signal detection in pharmacovigilance: empirical evaluation of data mining tools. *Pharmacoepidemiol Drug Saf.* Sep 2005;14(9):597-599.

**176.** Hauben M. A brief primer on automated signal detection. *Ann Pharmacother.* Jul-Aug 2003;37(7-8):1117-1123.

**177.** Hauben M. Early postmarketing drug safety surveillance: data mining points to consider. *Ann Pharmacother.* Oct 2004;38(10):1625-1630.

**178.** Hauben M, Reich L. Potential utility of data-mining algorithms for early detection of potentially fatal/disabling adverse drug reactions: a retrospective evaluation. *J Clin Pharmacol.* Apr 2005;45(4):378-384.

**179.** Hauben M, Reich L, DeMicco J, Kim K. 'Extreme duplication' in the US FDA Adverse Events Reporting System database. *Drug Saf.* 2007;30(6):551-554.

**180.** Hauben M, Reich L, Van Puijenbroek EP, Gerrits CM, Patadia VK. Data mining in pharmacovigilance: lessons from phantom ships. *Eur J Clin Pharmacol.* Nov 2006;62(11):967-970.

**181.** Hauben M, Zhou X. Quantitative methods in pharmacovigilance: focus on signal detection. *Drug Saf.* 2003;26(3):159-186.

**182.** Stephenson WP, Hauben M. Data mining for signals in spontaneous reporting databases: proceed with caution. *Pharmacoepidemiol Drug Saf.* Apr 2007;16(4):359-365.

**183.** Meyboom RH, Egberts AC, Edwards IR, Hekster YA, de Koning FH, Gribnau FW. Principles of signal detection in pharmacovigilance. *Drug Saf.* Jun 1997;16(6):355-365.

**184.** Budnitz DS, Pollock DA, Weidenbach KN, Mendelsohn AB, Schroeder TJ, Annest JL. National surveillance of emergency department visits for outpatient adverse drug events. *JAMA.* Oct 18 2006;296(15):1858-1866.

**185.** Budnitz DS, Shehab N, Kegler SR, Richards CL. Medication use leading to emergency department visits for adverse drug events in older adults. *Ann Intern Med.* Dec 4 2007;147(11):755-765.

**186.** Jhung MA, Budnitz DS, Mendelsohn AB, Weidenbach KN, Nelson TD, Pollock DA. Evaluation and overview of the National Electronic Injury Surveillance System-Cooperative Adverse Drug Event Surveillance Project (NEISS-CADES). *Med Care.* Oct 2007;45(10 Supl 2):S96-102.

**187.** Davis RL, Kolczak M, Lewis E, et al. Active surveillance of vaccine safety: a system to detect early signs of adverse events. *Epidemiology.* May 2005;16(3):336-341.

**188.** Kulldorff M, Davis R, Kolczak M. A maximized sequential probability ratio test for drug and vaccine safety surveillance.: Department of Ambulatory Care and Prevention.; 2007.

**189.** Curtis JR, Cheng H, Delzell E, et al. Adaptation of Bayesian data mining algorithms to longitudinal claims data: coxib safety as an example. *Med Care.* Sep 2008;46(9):969-975.

**190.** Norén G, Bate A, Hopstadius J, Star K, Edwards I. Temporal pattern discovery for trends and transient effects: its application to patient records. Paper presented at: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008; Las Vegas, Nevada, USA.

**191.** i3 Aperio. http://www.i3global.com/Solutions/DrugSafety/i3Aperio/. Accessed 4 June 2010.

**192.** Dore DD, Seeger JD, Arnold Chan K. Use of a claims-based active drug safety surveillance system to assess the risk of acute pancreatitis with exenatide or sitagliptin compared to metformin or glyburide. *Curr Med Res Opin.* Apr 2009;25(4):1019-1027.

**193.** Beach KJ, Le HV, Powell G, Pattishall E, Ryan PB, Mera R. Performance of a Semi-Automated Method for Risk Estimation using Observational Databases; 2009.

**194.** Merrill GH, Ryan, P.B., Painter, J.L. Using SNOMED to Normalize and Aggregate Drug References in the SafetyWorks Observational Pharmacovigilance Project. *KR-MED*. Phoenix, AZ, USA; 2008.

**195.** Ryan PB, Powell, G.E. Exploring Candidate Differences Between Drug Cohorts Prior To Exposure: A Systematic Approach Using Multiple Observational Databases. *International Society of Pharmacoeconomics and Outcomes Research*. Toronto, ON, CAN; 2008.

**196.** Ryan PB, Painter, J.L., Merrill, G.H. Defining medical conditions by mapping ICD-9 to MedDRA: A systematic approach to integrating disparate observational data sources for enabling enhanced pharmacovigilance analyses. *Drug Information Association*. Boston, MA, USA; 2008.

**197.** FDA. FDA Awards Contract to Harvard Pilgrim to Develop Pilot for Safety Monitoring System. http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm196968.htm. Accessed 4 June 2010.

**198.** EU-ADR. http://www.euadr-project.org/drupal/?q=home. Accessed 4 June 2010.

**199.** Trifiro G, Fourrier-Reglat A, Sturkenboom MC, Diaz Acedo C, Van Der Lei J. The EU-ADR project: preliminary results and perspective. *Stud Health Technol Inform.* 2009;148:43-49.

**200.** Avillach P, Mougin F, Joubert M, et al. A semantic approach for the homogeneous identification of events in eight patient databases: a contribution to the European eu-ADR project. *Stud Health Technol Inform.* 2009;150:190-194.

**201.** Trifiro G, Pariente A, Coloma PM, et al. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoepidemiol Drug Saf.* Dec 2009;18(12):1176-1184.

**202.** IMI-PROTECT. http://www.imi-protect.eu/. Accessed 4 June 2010.

**203.** Observational Medical Outcomes Partnership. http://omop.fnih.org. Accessed November 28, 2009.

**204.** Platt R, Madre L, Reynolds R, Tilson H. Active drug safety surveillance: a tool to improve public health. *Pharmacoepidemiol Drug Saf.* Dec 2008;17(12):1175-1182.

**205.** Platt R, Wilson M, Chan KA, Benner JS, Marchibroda J, McClellan M. The new Sentinel Network--improving the evidence of medical-product safety. *N Engl J Med.* Aug 13 2009;361(7):645-647.

**206.** Sturmer T, Rothman KJ, Avorn J. Pharmacoepidemiology and "in silico" drug evaluation: is there common ground? *J Clin Epidemiol.* Mar 2008;61(3):205-206.

**207.** Rothman KJ, Poole C. A strengthening programme for weak associations. *Int J Epidemiol.* Dec 1988;17(4):955-959.

**208.** Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Epidemiology.* Nov 2007;18(6):805-835.

**209.** Guidelines for good pharmacoepidemiology practices (GPP). *Pharmacoepidemiol Drug Saf.* Feb 2008;17(2):200-208.

**210.** Sturmer T, Carey T, Poole C. ISPOR Health Policy Council Proposed Good Research Practices for Comparative Effectiveness Research: Benefit or Harm? *Value Health.* Oct 8 2009.

**211.** Ryan PB, Griffin D, Reich C, et al. OMOP Common Data Model (CDM) Specifications. http://omop.fnih.org/CDMandTerminologies. Accessed 30 May 2010.

**212.** D'Agostino RB, Jr., D'Agostino RB, Sr. Estimating treatment effects using observational data. *JAMA.* Jan 17 2007;297(3):314-316.

**213.** Rubin DB. On principles for modeling propensity scores in medical research. *Pharmacoepidemiol Drug Saf.* Dec 2004;13(12):855-857.

**214.** Kurth T, Walker AM, Glynn RJ, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol.* Feb 1 2006;163(3):262-270.

**215.** Sturmer T, Schneeweiss S, Brookhart MA, Rothman KJ, Avorn J, Glynn RJ. Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. *Am J Epidemiol.* May 1 2005;161(9):891-898.

**216.** McMahon AD, Evans JM, McGilchrist MM, McDevitt DG, MacDonald TM. Drug exposure risk windows and unexposed comparator groups for cohort studies in pharmacoepidemiology. *Pharmacoepidemiol Drug Saf.* Jul 1998;7(4):275-280.

**217.** Ryan PB, Griffin, D., Whittenburg, L., Foltz, D., Overhage, J.M. Points to Consider in Developing a Common Semantic Data Model and Terminology Dictionary for Observational Analyses. http://omop.fnih.org/CDMandTerminologies22May2009. Accessed November 28, 2009.

**218.** Gross R, Bilker WB, Strom BL, Hennessy S. Validity and comparison of two measures of days supply in Medicaid claims data. *Pharmacoepidemiol Drug Saf.* Oct 2008;17(10):1029-1032.

**219.** Ryan PB. Establishing a Condition Era Persistence Window for Active Surveillance. http://omop.fnih.org/OMOPWhitePapers. Accessed March 8, 2011.

**220.** Norris S, Weinstein J, Peterson K, Thakurta S. *Drug class review: Direct renin inhibitors, angiotensin converting enzyme inhibitors, and angiotensin II receptor blockers.* 2010.

**221.** Heran BS, Wong MM, Heran IK, Wright JM. Blood pressure lowering efficacy of angiotensin converting enzyme (ACE) inhibitors for primary hypertension. *Cochrane Database Syst Rev.* 2008(4):CD003823.

**222.** Matchar DB, McCrory DC, Orlando LA, et al. Systematic review: comparative effectiveness of angiotensin-converting enzyme inhibitors and angiotensin II receptor blockers for treating essential hypertension. *Ann Intern Med.* Jan 1 2008;148(1):16-29.

**223.** Chobanian A. *The joint national committee on prevention detection and evaluation of high blood pressure.* Bethesda, MD: US department of health and human services NHLBI; 2009.

**224.** Standards of medical care in diabetes--2010. *Diabetes Care.* Jan;33 Suppl 1:S11-61.

**225.** K/DOQI clinical practice guidelines for bone metabolism and disease in chronic kidney disease. *Am J Kidney Dis.* Oct 2003;42(4 Suppl 3):S1-201.

**226.** Agusti A, Bonet S, Arnau JM, Vidal X, Laporte JR. Adverse effects of ACE inhibitors in patients with chronic heart failure and/or ventricular dysfunction : meta-analysis of randomised clinical trials. *Drug Saf.* 2003;26(12):895-908.

**227.** McDowell SE, Coleman JJ, Ferner RE. Systematic review and meta-analysis of ethnic differences in risks of adverse reactions to drugs used in cardiovascular medicine. *BMJ.* May 20 2006;332(7551):1177-1181.

**228.** Lakhdar R, Al-Mallah MH, Lanfear DE. Safety and tolerability of angiotensin-converting enzyme inhibitor versus the combination of angiotensin-converting enzyme inhibitor and angiotensin receptor blocker in patients with left ventricular dysfunction: a systematic review and meta-analysis of randomized controlled trials. *J Card Fail.* Apr 2008;14(3):181-188.

**229.** Kostis JB, Shelton B, Gosselin G, et al. Adverse effects of enalapril in the Studies of Left Ventricular Dysfunction (SOLVD). SOLVD Investigators. *Am Heart J.* Feb 1996;131(2):350-355.

**230.** 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). *JAMA.* Dec 18 2002;288(23):2981-2997.

**231.** Yusuf S, Sleight P, Pogue J, Bosch J, Davies R, Dagenais G. Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. The Heart Outcomes Prevention Evaluation Study Investigators. *N Engl J Med.* Jan 20 2000;342(3):145-153.

**232.** Chalmers D, Whitehead A, Lawson DH. Postmarketing surveillance of captopril for hypertension. *Br J Clin Pharmacol.* Sep 1992;34(3):215-223.

**233.** Speirs C, Wagniart F, Poggi L. Perindopril postmarketing surveillance: a 12 month study in 47,351 hypertensive patients. *Br J Clin Pharmacol.* Jul 1998;46(1):63-70.

**234.** Brown NJ, Ray WA, Snowden M, Griffin MR. Black Americans have an increased rate of angiotensin converting enzyme inhibitor-associated angioedema. *Clin Pharmacol Ther.* Jul 1996;60(1):8-13.

**235.** Messner Pellenc P, Rudnicki A, Leclercq F, Grolleau R. Enalapril in the treatment of mild-to-moderate heart failure in general medical practice: a prospective and multicentre study concerning 17,546 patients. *Acta Cardiol.* 1995;50(3):187-201.

**236.** Palmer BF. Managing hyperkalemia caused by inhibitors of the renin-angiotensin-aldosterone system. *N Engl J Med.* Aug 5 2004;351(6):585-592.

**237.** Thorp ML, Ditmer DG, Nash MK, et al. A study of the prevalence of significant increases in serum creatinine following angiotension-converting enzyme inhibitor administration. *J Hum Hypertens.* May 2005;19(5):389-392.

**238.** DailyMed. http://dailymed.nlm.nih.gov/dailymed/about.cfm. Accessed 4 June 2010.

**239.** Whitaker H. The self controlled case series method. *BMJ.* 2008;337:a1069.

**240.** Hasselblad V, McCrory DC. Meta-analytic tools for medical decision making: a practical guide. *Med Decis Making.* Jan-Mar 1995;15(1):81-96.

**241.** Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* Sep 6 2003;327(7414):557-560.

**242.** *Report to the President Realizing the Full Potential of Health Information Technology to Improve Healthcare for Americans: The Path Forward*: President's Council of Advisors on Science and Technology; 2010.

**243.** Woodcock J, Behrman RE, Dal Pan GJ. Role of postmarketing surveillance in contemporary medicine. *Annu Rev Med.* Feb 18 2011;62:1-10.

**244.** Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System - A National Resource for Evidence Development. *N Engl J Med.* Jan 12 2011.

**245.** Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med.* Nov 2 2010;153(9):600-606.

**246.** IMI-PROTECT. http://www.imi-protect.eu/. Accessed November 28, 2009.

**247.** Coloma PM, Schuemie MJ, Trifiro G, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiol Drug Saf.* Jan 2011;20(1):1-11.

**248.** *Federal Coordinating Council for Comparative Effectiveness Research Report to the President and the Congress*: US Department of Health and Human Services; 2009.

**249.** Lieu TA. *Data Needs for Signal Refinement: Experience from the HMO Research Network and Vaccine Safety Datalink Project* 2010.

**250.** Selby JV, Fireman B, Butler M. *Report to FDA on a Protocol for Active Surveillance of Acute Myocardial Infarction in Association With Use of a Pharmaceutical Agent* 2010.