EVALUATION OF THE PERFORMANCE OF THE HIGH-DIMENSIONAL PROPENSITY SCORE ALGORITHM TO ADJUST FOR CONFOUNDING OF TREATMENT EFFECTS ESTIMATED IN HEALTHCARE CLAIMS DATA

Hoa Van Le, MD

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 Department of Epidemiology in the Gillings School of Global Public Health.

Chapel Hill
2012

Approved by:

Til Stürmer, MD, PhD

Charles Poole, ScD

M. Alan Brookhart, PhD

Victor J. Schoenbach, PhD

Kathleen J. Beach, MD, MPH

ABSTRACT

HOA VAN LE: Evaluation of the Performance of the High-Dimensional Propensity Score Algorithm to Adjust for Confounding of Treatment Effects Estimated in Healthcare Claims Data
(Under the direction of Professor Til Stürmer)

The aim of the high-dimensional propensity score (hd-PS) algorithm is to select and adjust for baseline confounders in pharmacoepidemiologic studies based on healthcare claims data. It is not well understood how the performance of the hd-PS is affected by 1) the channelling of drugs at specific calendar time periods and differences in administrative claims databases; 2) low outcome incidence or exposure prevalence in medium sized or large cohorts; and 3) aggregation of medical diagnoses and medications in cohorts with small size, low outcome incidence and low exposure prevalence.

We estimated risk ratios for upper gastrointestinal complication in patients with rheumatoid arthritis or osteoarthritis after initiating oral celecoxib versus ibuprofen or diclofenac in two large longitudinal healthcare claims databases. We conducted separate analyses for subcohorts before and after withdrawal of rofecoxib, a drug in the same class as celecoxib. We applied the hd-PS algorithm using a combination of demographic, predefined and hd-PS covariates with either PS deciles or 1:1 greedy matching for each cohort. In addition, we conducted pooled analyses for two combined databases stratified by data source and adjusted by either deciles of separate PSs or 1:1 greedy matching within the data source. The different methods of propensity score confounder selection inconsistently reduced confounding by indication across calendar time periods and administrative data sources.

To evaluate the effects of aggregation of medical diagnoses and medications on the performance of the hd-PS, we resampled studies to assess the influence of size, outcome incidence,

and exposure prevalence. For each sample, baseline covariates were identified with and without the hd-PS algorithm to estimate the treatment effect using propensity score deciles. In an empirical pharmacoepidemiologic study using claims data, aggregations of medications into chemical, pharmacological or therapeutic subgroups (level 4) of the Anatomical Therapeutic Chemical classification alone or in combination of aggregation of diagnoses into largest groups (level 1) of the Clinical Classification Software improved the hd-PS adjustment for confounding in most scenarios including ones with small cohort size, rare outcome incidence, and low exposure prevalence.

ACKNOWLEDGEMENTS

I am greatly indebted to Dr. Til Stürmer and Dr. Victor J. Schoenbach for their tremendous advice and support, and great patience specifically during the dissertation process and study period at UNC.

I would like to express my gratitude to the members of my dissertation committee: Dr.Charles Poole, Dr. M. Alan Brookhart and Dr. Kathleen J. Beach for their many invaluable comments and advice.

I would like to thank to all of my dear teachers, classmates, friends, and colleagues at UNC and GSK for their wonderful support.

I would like to acknowledge the Vietnam Education Foundation, UNC Department of Epidemiology, GlaxoSmithKline, Harry Guess-Merck Scholarship, UNC Graduate School and International Society for Pharmacoepidemiology for financial support.

Last but not least, I want to say thank you to my parents Thang Van Le and Nong Thi Nguyen, my parents in laws Diep Van Truong and Le Thi Nguyen, my lovely wife Chi Thi Le Truong, my three cute daughters Chi Huu Hong Le, Phuong Huu Uyen Le and Ngan Huu Kim Le, my brothers and sisters. I am sure that they are all proud of my achievement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AHRQ	Agency for Healthcare Research and Quality
ATC	Anatomical Therapeutic Chemical (ATC) classification
CLASS	Celecoxib Long-term Safety Study
CCS	Clinical Classifications Software
EHR	Electronic Health Records
GPRD	General Practice Research Database
hd-PS	High-Dimensional Propensity Score
HIPAA	Health Insurance Portability and Accountability Act
HR	Hazard Ratio
ICD-9	International Statistical Classification of Diseases 9th version
IRR	Incidence Rate Ratio
NSAIDs	Nonsteroidal Anti-inflammatory Drugs
tNSAIDs	Traditional Nonsteroidal Anti-inflammatory Drugs
OA	Osteoarthritis
OR	Odds Ratio
PS	Propensity Score
RA	Rheumatoid Arthritis
RCT	Randomized Controlled Trial
RR	Risk Ratio
UGI	Upper Gastrointestinal
SNOMED CT	Systematized Nomenclature of Medicine-Clinical Terms

CHAPTER I

STATEMENT OF SPECIFIC AIMS

Although early detection and assessment of drug safety signals is important [1-3], post-marketing drug safety studies often face challenges such as small size, rare incidence of adverse outcomes, or low exposure prevalence after a new drug launch. In addition, active surveillance will often generate a large number of safety signals, which emphasizes the need for a method that can rapidly, yet systematically, refine a signal. Large healthcare claims databases are important sources for active surveillance. However, there is potential channelling bias of drugs, different patients, different providers, healthcare plans, and payers over time in disparate administrative healthcare databases.

Propensity score methods are an increasingly used approach to control for measured potential confounders, especially in pharmacoepidemiologic studies of rare outcomes in the presence of many covariates from different data dimensions encountered in administrative healthcare databases [4-7]. Methods of selecting variables for propensity score models based on substantive knowledge have been proposed [8-12]. However, substantive knowledge may often be lacking, and identification of a very large pool of potential confounders for propensity score model is still a major challenge. The High-Dimensional Propensity Score (hd-PS) algorithm automatically defines and selects variables for inclusion in the propensity score to adjust treatment effect estimates in studies using healthcare data [13, 14]. The hd-PS seems interesting as it leads to confounding control that is as least as good as the one obtained by adjustment limited to covariates predefined by expert knowledge [13-16]. The hd-PS algorithm could reduce programming time and error and run in studies pooling multiple claims databases [13, 14] and its performance has been evaluated with few outcome events or few exposed subjects in small cohorts [17]. It is a promising algorithm for studies using

healthcare claims data. However, it is not known whether different calendar time periods, data sources, low outcome incidence or exposure prevalence can degrade hd-PS performance in medium sized or large cohorts. Also, no study to date has assessed how hd-PS performance is affected by aggregating medical diagnoses and/or medications, especially in cohorts with relatively few patients, rare outcome incidence, or low exposure prevalence. Extensive testing of the performance of hd-PS should be carried out in multiple settings to provide more confidence and to determine the value of this new approach.

This dissertation addresses the following questions:

1) How do different calendar time periods or administrative data sources affect the performance of the hd-PS?
2) How does low outcome incidence or exposure prevalence degrade hd-PS performance in medium sized or large cohorts?
3) How does aggregating medical diagnoses and/or medications affect the hd-PS performance, especially in cohorts with relatively few patients, rare outcome incidence, or low exposure prevalence?

To answer these questions, the following specific aims are addressed in this research:

Specific aim 1: To evaluate the performance of the hd-PS algorithm to adjust for confounding of treatment effects in cohorts with different calendar time periods and administrative data sources.

Hypothesis for specific aim 1: As we use an established association of upper gastrointestinal complications with celecoxib versus traditional non-steroidal anti-inflammatory agents (tNSAIDs) in rheumatoid arthritis (RA) or osteoarthritis (OA) patients [18-23] to evaluate the hd-PS performance and on 30 September 2004 Merck Inc. announced the voluntary withdrawal of rofecoxib from the US market [24]. We hypothesized that channelling of celecoxib, a drug of the same class as rofecoxib, would be affected by the withdrawal of rofecoxib. Specifically, we hypothesized that the influence of

upper gastrointestinal (UGI) risk on preferential prescribing of coxibs would increase (increased confounding by indication). The channelling of celecoxib at different calendar time periods, or differences in administrative claims databases will affect the hd-PS performance.

Rationale for specific aim 1: As an automated procedure, the hd-PS does not have options for users to consider specific subtleties of the data. These might include, but are not restricted to, expected changes in the channelling of drugs at specific calendar time points (e.g., due to publication of landmark trials, black box warnings, 'Dear Doctor' letters, marketing activities by drug makers, new guidelines or policies, etc.) [25], differences in study periods, and differences in database sources. However, it is not known whether the hd-PS performs well in these situations. An assessment of the hd-PS performance may provide insight into a guidance of using hd-PS for different calendar time periods and disparate data sources.

Specific aim 2: To determine how low outcome incidence or exposure prevalence can degrade hd-PS performance in medium sized or large cohorts.

Hypothesis for specific aim 2: The performance of the hd-PS will degrade not only in cohorts with small size, but also in cohorts with low outcome incidence or infrequent exposure prevalence.

Rationale for specific aim 2: The hd-PS algorithm prioritizes variables by their potential for confounding control based on their prevalence and on bivariate associations of each covariate with the treatment and with the study outcome [13, 26]. In cohorts with either lower outcome incidence or exposure prevalence, there is a higher number of baseline confounders (e.g., those with low prevalence, missing covariate-exposure association, zero/undefined covariate-outcome association) not meeting hd-PS inclusion criteria. The hd-PS performance has been evaluated with few outcome events or few exposed subjects in small cohorts only [17]. Updated information will provide evidence on the effects of low outcome incidence or exposure prevalence on the hd-PS performance in medium sized or large cohorts.

Specific aim 3: To evaluate the effects of aggregating medical diagnoses and/or medications on the hd-PS performance, especially in cohorts with relatively few patients, rare outcome incidence, or low exposure prevalence.

Hypothesis for specific aim 3: Aggregation may improve control of confounding, by increasing the prevalence of rare baseline potential confounders so they would be included in the hd-PS, or may worsen control of confounding, by increasing the amount of information bias in control variables.

Rationale for specific aim 3: The hd-PS algorithm prioritizes variables by their potential for confounding control based on their prevalence and on bivariate associations of each covariate with the treatment and with the study outcome [13, 26]. Combining medical diagnoses or medications into higher-level categories reduces the number of baseline potential confounders (e.g., those with low prevalence, missing covariate-exposure association, zero/undefined covariate-outcome association) not meeting hd-PS inclusion criteria. No study to date has assessed how hd-PS performance is affected by aggregating medical diagnoses and/or medications, especially in cohorts with relatively few patients, rare outcome incidence, or low exposure prevalence. Updated information will provide evidence on the effects of aggregating medical diagnoses and/or medications on hd-PS performance, especially in cohorts with relatively few patients, rare outcome incidence, or low exposure prevalence.

CHAPTER II

LITERATURE REVIEW

A. BACKGROUND

Significance of the High-Dimensional Propensity Score algorithm

As the passive drug safety surveillance system has well-recognized drawbacks [27-30], active safety surveillance such as the Sentinel System, a part of the Food and Drug Administration's Sentinel Initiative, using patient information derived from health insurance claims data [31] is one of the basic methods of signal detection that should be developed [32-34]. However, the active surveillance will most likely generate a large number of safety signals, which emphasizes the need for an approach that can earlier [1-3, 35,36] and more rapidly, yet systematically, refine a signal [36-39]. One approach might be to automate the assessment of the relation between a drug exposure and a medical condition with an ability to conduct evaluations in disparate electronic healthcare claims or medical record databases [40,41]. The hd-PS with an automated process for adjustment of a large number of candidate covariates for propensity score model could significantly contribute to the early refinement of drug safety signals [13].

Proxy adjustment for patient health status in longitudinal health claims data

Several levels of proxies for the health state of patients are present in the longitudinal healthcare claims data through drug dispensing, medical diagnoses, procedures, providers, and health insurance plans. Healthcare claims databases have incomplete data on potential confounders such as disease severity, over-the-counter medication use, race/ethnicity, smoking status, body mass index, laboratory results (and their laboratory specific normal ranges), cognitive and functional status, and socio-economic status. These factors are potential confounders of an association between drugs

and outcomes in pharmacoepidemiologic studies. Chains of proxies might be surrogates for access to care [42], condition severity, physician ability, drug preference [43] or medication co-payment ability [43]. Seeger et al. proposed that health care claims may serve as proxies, in hard-to-predict ways, for important unmeasured variables [45]. Following this general idea of controlling for a wide variety of covariates many of which would not be readily seen as confounders outside of studies based on healthcare claims data, Stürmer et al. used propensity score models with over 70 variables representing medical codes present during a baseline period [5]. Johannes et al. created a propensity score model that considered as candidate variables the 100 most frequently occurring diagnoses, procedures, and outpatient medications in healthcare claims [46].

Variable selection for propensity score model

Propensity score methods as formalized by Rosenbaum and Rubin [47] are an increasingly used approach to control for measured potential confounders, especially in pharmacoepidemiologic studies of rare outcomes in the presence of many covariates from different data dimensions encountered in administrative healthcare databases [4-7]. Propensity scores can be implemented by stratification, matching, weighting, or as a continuous covariate in the outcome model [48-51]. The use of propensity score models with many covariates on observational assessment of treatment effects was explored in several studies [45,52]. Addition of clinical covariates into propensity score models may lead to better control of confounding than models with less covariate information in specific settings [53,54] but to identify a very large pool of potential confounders for propensity score models is still a major challenge. Methods of selecting variables for propensity score models based on substantive knowledge have been proposed [8-12]. Brookhart et al. demonstrated that including variables in the propensity score model that are associated with the exposure but not the outcome will increase variance of the estimator with no improvement in confounding control, and may introduce confounding in some situations [11]. Moreover, substantive knowledge may often be lacking, and the meaning of various medical codes may often be unclear [55]. In a nutshell the major challenge is to include a sufficient set of confounders and risk factors for the outcome of interest in the model while

avoiding the inclusion of instrumental variables [11,56], variables affected by the treatment (or the outcome), and colliders in M-structures [12].

The high dimensional propensity score algorithm

A recently-developed strategy for selecting from a large pool of baseline covariates for propensity score analyses is the use of a computer-applied algorithm [13,57], such as the High-Dimensional Propensity Score (hd-PS) algorithm. The hd-PS automatically defines and selects variables for inclusion in the propensity score to adjust treatment effect estimates in studies using healthcare claims data [13,14]. The hd-PS macro [13] is a multi-step algorithm to implement proxy adjustment in claims data. The macro's steps include:

1) identify data dimensions: the hd-PS uses the health service records of patients e.g., 5 common data dimensions: pharmacy claims, outpatient diagnoses, outpatient procedures, inpatient diagnoses, and inpatient procedures;

2) identify candidate covariates: because the prevalence of a binary factor is symmetrical around 0.5, the hd-PS subtracts all prevalence estimates larger than 0.5 from 1.0. In addition to basic variables e.g. age, gender, race, and calendar time, the hd-PS identifies most prevalent covariates from each data dimension (e.g., top n=200);

3) assess recurrence of same code: each code is assessed for within-patient occurrence during a predefined period (e.g., 6 months) and divided into three binary variables: once, sporadic ≥ median number of times, or frequent ≥ 75^{th} percentile number of times. A code would have a "true" value for all three recurrence variables if it occurred above the 75^{th} percentile number of times. If any of the values were equal, the hd-PS dropped the variable with the higher cutpoint;

4) prioritize covariates: the hd-PS algorithm drops covariates with fewer than 100 patients (exposed and unexposed combined) per variable, with missing ("zero/undefined") covariate-exposure association, and with missing ("zero/undefined") covariate-outcome association from the prioritization. The covariate-exposure association is missing when the prevalence of covariate in either exposed

group or unexposed group is zero or 1. The covariate-outcome association is missing when the prevalence of covariate in cases or noncases is zero or 1. The hd-PS also drops covariate if the natural log of its multiplicative bias is missing. For example, covariates with missing covariate-outcome or covariate-exposure associations, with the same symmetric prevalence in exposed and unexposed group, are dropped because they lead the confounding effect equal the null or undefined when we insert them into the Bross formula [13, 26]. Remained variables from data dimensions are prioritized by their potential for confounding control based on the symmetric prevalence of each covariate in the treated and untreated and the bivariate association of the covariate with the study outcome based on absolute value of natural log of multiplicative bias derived by Bross formula after adjusting for demographic covariates [13, 26].

5) select covariates for adjustment: by default, the top k=500 indicator variables from step 4 are selected in addition to age, gender, race and calendar year;

6) estimate exposure propensity score: a propensity score is estimated for each subject as a predicted probability of exposure conditional on all covariates at step 5 using multivariate logistic regression;

7) estimate propensity score-adjusted outcome models: the algorithm groups subjects into propensity score deciles and uses multivariate regression analyses to model the study outcome as a function of exposure and indicator terms for propensity score deciles.

The hd-PS led to adjustment for confounding that was at least as good as the one obtained from variable selection based on expert knowledge in a few selected examples. There is even some very limited evidence that it leads to confounding control that is as least as good as the one obtained by adjustment limited to covariates predefined by expert knowledge [13,15,16]. In these studies [13,15,16] the estimated risk reduction of UGI complication after adjustment for investigator-specified covariates and the hd-PS algorithm was 6-21% and 12-22%. Approximately 50% lower risk of UGI complication among coxib initiators compared with tNSAID initiators was reported in RCT finding [18-23]. We therefore assume that a treatment effect estimate closer to 0.5 is less biased by confounding.

The hd-PS algorithm could reduce programming time and error, and run in studies pooling multiple claims databases [13,14]. Prior studies demonstrated the hd-PS was potential algorithm software for active drug safety monitoring systems using longitudinal healthcare claims databases [14]. However, extensive testing of the performance of hd-PS should be carried out in multiple settings to provide more confidence and to determine the value of this new approach. Moreover, any solutions to improve the hd-PS performance particularly in specific settings will contribute to the research community for active drug safety surveillance.

B. FACTORS AFFECTING PERFORMANCE OF HIGH-DIMENSIONAL PROPENSITY SCORE ALGORITHM

Changes of channelling bias and calendar time periods

Channeling bias (here defined as confounding by indication) is a serious threat to the validity of nonexperimental studies of treatment effects [58,59]. Walker et al. defines confounding by indication as a resul of differential selection of patients into drug exposure groups [60]. Schneeweiss et al. reported patient-, provider-, and system-related factors caused potential bias due to channeling of patients to the newly marketed medication [58]. Factors influencing the prescription of drugs by physicians can vary by physician, change over time [61], and are often based on patients' characteristics, diagnoses, prognoses, and behaviors. These factors are usually not directly recorded in administrative databases [58,59]. A failure to adjust for imbalanced factors that are predictors of the outcome of interest leads to biased results. Publications raised the importance of the channeling bias, especially for new drugs after market launch [58]. There is still a lack of guidance on how to deal with this challenge. It is unknown if strategies exist to separate which of these factors are potential confounders or instrumental variables [59].

On 30 September 2004 Merck Inc. announced the voluntary withdrawal of rofecoxib from the US market [24]. As an automated procedure, however, the hd-PS does not consider specific subtleties of the data. These might include, but are not restricted to, expected changes in the channelling of drugs at specific calendar time points (e.g., due to publication of landmark trials, black box warnings, 'Dear Doctor' letters, marketing activities by drug makers, new guidelines or policies,

etc.) [25,52], and differences in study periods and administrative healthcare databases. Prior research studies demonstrated the challenges of channelling bias in nonrandomized studies, but no studies addressed how the changes of channelling bias due to the withdrawal of a drug in the same class can affect hd-PS performance.

Different data sources and pooled analyses

There has been increased interest in using automated healthcare claims databases as a useful data source for researchers and regulatory agencies to study the safety of drugs, particularly for rare outcomes in post-marketing studies. The administrative data have some advantages e.g., longitudinal data, accuracy in recording the date of dispensing and less biased by knowledge of the study outcome, representativeness of routine clinical practice in large populations. However, these data were generated primarily for administrative purposes and have disadvantages. The claims databases do not completely capture all of a patients' medical history (e.g., incomplete capture of healthcare or treatments provided outside of health plan coverage, alcohol use, inpatient drugs, over-the-counter medications, medications that cost less than co-payments, dispensed medication less than consumed medication). Many factors are involved in the data generation process and in the creation of quality for a specific database. These factors include coding errors, under-reporting of secondary diagnoses, changes in hardware, software, or coding practice over time, and mergers of healthcare plans leading to doubling/sharing patient identification. Each database has its own "specific way" to generate data, and this is often undocumented or not updated [59]. Hennessy et al. reported that descriptive analyses of the population composition over time can help one determine the integrity of linked administrative databases [62]. The literature often compared characteristics of electronic medical records and claims databases. Each claims database is often promoted with its own features and benefits. There is little information to directly compare the quality of "similar" of claims databases.

Pharmacoepidemiologic studies usually need large databases pooled from many administrative data sources. Pooled analyses from multiple populations have advantages for rare

outcome [1-4]. Recently, Rassen et al. developed a novel method of pooled analyses to use hd-PS which delivered similar point estimates for multi-center studies [63]. The PS-based pooling method [63] using separate propensity scores estimated from each data source, showed some benefits for the study of the same drug-outcome pair in multiple databases. The pooled analyses were stratified by the center and adjusted by deciles of separate PSs [63] where the PS distribution without trimming of the non-overlap region to create separate PS deciles for which each data source cohort was employed [63]. In addition to pooled analyses using deciles of separate PSs, the pooled analyses can be carried in the matched cohorts after 1:1 PS greedy matching [64] starting at the eighth decimal place [63] and continued to the first within the study center.

Prior research studied the performance of the hd-PS algorithm with established drug-outcomes in US healthcare claims and UK electronic medical record databases [13,15,16]. There are two large longitudinal administrative US databases: MarketScan® commercial claims and encounters of Thomson Reuters healthcare [65] and Optum® Impact® National Managed Care Benchmark Database [66]. MarketScan is a longitudinal 10-year healthcare claims database which captures patient demographics, inpatient and outpatient diagnoses and procedures, and medications from a selection of large private employers, health plans, government agencies and other public organizations. Optum is a longitudinal medical claims history for more than 98 million individuals. The Optum data come from more than 46 health plans in the US with available information: patient demographics such as age and gender, diagnoses, procedures, and medications recorded during outpatient visits and hospital admissions. These two large healthcare claims databases are potential candidates to evaluate the performance of the hd-PS where proxies for the health state of patients are present in the longitudinal health claims data through drug dispensing, medical diagnosis and procedure. The chains of proxies can be surrogates for access to care [42] condition severity, physician ability, drug preference [43] or medication co-payment ability [44]. There was no study to evaluate how the different data sources can affect the hd-PS performance.

The PS-based pooling method [63] was developed for pooling analyses of multiple studies. There was no study to evaluate whether similar benefits of the pooled and individual data analyses

with propensity score deciles and greedy matching could be observed in studies using two claims data sources, particularly in pre and post-withdrawal subcohorts with potential changes of channelling bias due to rofecoxib withdrawal.

Small samples, rare outcome incidence and low exposure prevalence

Because early detection and assessment of drug safety signals is very important [1-3,37] there is a possible delay if the hd-PS can perform in a large sample size only. The hd-PS algorithm prioritizes variables by their potential for confounding control based on their prevalence and on bivariate associations of each covariate with the treatment and with the study outcome [13,26]. Rassen et al. reported that hd-PS functioned well in small cohorts with >50 exposed patients with an outcome event; and using zero-cell correction or exposure-based covariate selection permitted hd-PS to function robustly with 25–50 exposed patients with an outcome event and to yield estimates closer to estimates obtained in the full cohort [17]. The prior study concluded that few exposed events and few exposed subjects affected the performance of the hd-PS in small samples. In reality, few exposed events will tend to be the norm after a new drug is launched to the market (low exposure prevalence) or due to rare events (low outcome incidence). There is no study so far that has evaluated how few exposed events or few exposed subjects can affect performance of the hd-PS in medium sized and large samples.

The hd-PS algorithm prioritizes variables by their potential for confounding control based on their prevalence and on bivariate associations of each covariate with the treatment and with the study outcome [13,26]. In cohorts with either lower outcome incidence or exposure prevalence, there will be a larger number of baseline potential confounders (e.g., those with low prevalence, missing covariate-exposure association, zero/undefined covariate-outcome association) not meeting hd-PS inclusion criteria.

C. MEDICAL CODING AND AGGREGATIONS

Major U.S. administrative databases represent medical diagnoses using International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9) codes. ICD-9 does not rely on

its hierarchical relationships [67], but the Clinical Classification Software (CCS) developed by the Agency for Healthcare Research and Quality (AHRQ) can be used to group diagnoses into clinically meaningful categories [68]. Similarly, medications can be grouped into levels of the Anatomical Therapeutic Chemical (ATC) classification [69].

Clinical Classification Software (CCS)

CCS is a tool [68] for clustering patient diagnoses and procedures into clinically meaningful categories. These categories were used as covariates in several studies [70,71]. CCS collapses diagnosis codes from ICD-9, which contains more than 13,600 diagnosis codes divided into 18 categories [68]. CCS is unique as a grouping approach because it does not mix diagnoses with treatment. There is available cross-mapping aggregation from ICD-9 medical diagnoses to CCS with frequent maintenance and update at AHRQ [68]. There are 18, 134, 355 and 207 categories in CCS levels 1, 2, 3 and 4, respectively [68]. Examples of aggregations of ICD-9 diagnosis codes into CCS levels are in Figure 2.1 and Appendix A.

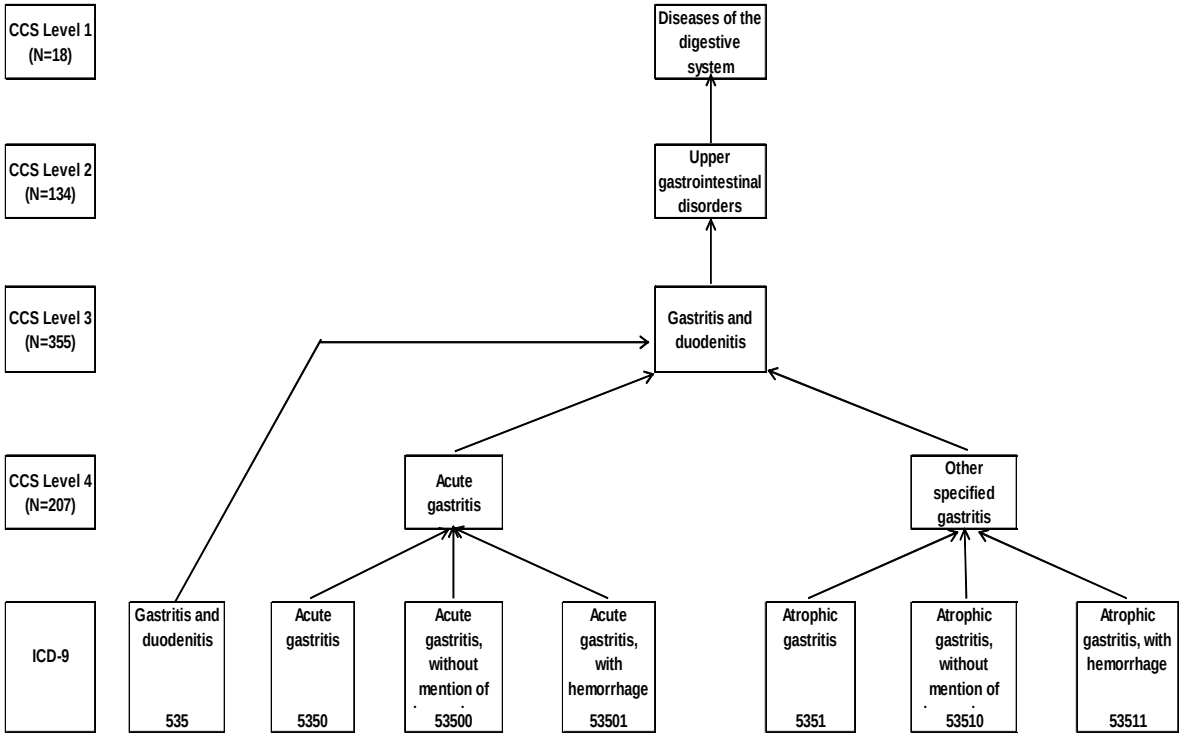

Figure 2.1. An example of aggregations of ICD-9 diagnosis codes into 4 levels of the Clinical Classification Software

Anatomical Therapeutic Chemical (ATC) classification

The Anatomical Therapeutic Chemical (ATC) classification of the World Health Organization (WHO) [69] classifies active substances into different groups based on their target organ or system and their therapeutic, pharmacological and chemical properties. Drugs are classified into fourteen main groups (1st level) with pharmacological or therapeutic subgroups (2nd level). The 3rd and 4th levels are chemical, pharmacological or therapeutic subgroups, and the 5th level is the chemical substance. Several ATC groups are subdivided into both chemical and pharmacological groups. The pharmacological group is often chosen if a new substance fits in both a chemical and pharmacological 4th level. Substances in the same 4th ATC level are not pharmacotherapeutically equivalent, as they may have different modes of action, therapeutic effects, drug interactions and adverse drug reaction profiles. New 4th levels are commonly established if at least two approved substances fit in the group. A new substance not clearly belonging to any existing group of related

substances of ATC 4th level will often be placed in an X group ("other" group) [69]. An example of aggregations of medications into ATC levels is in Figure 2.2.

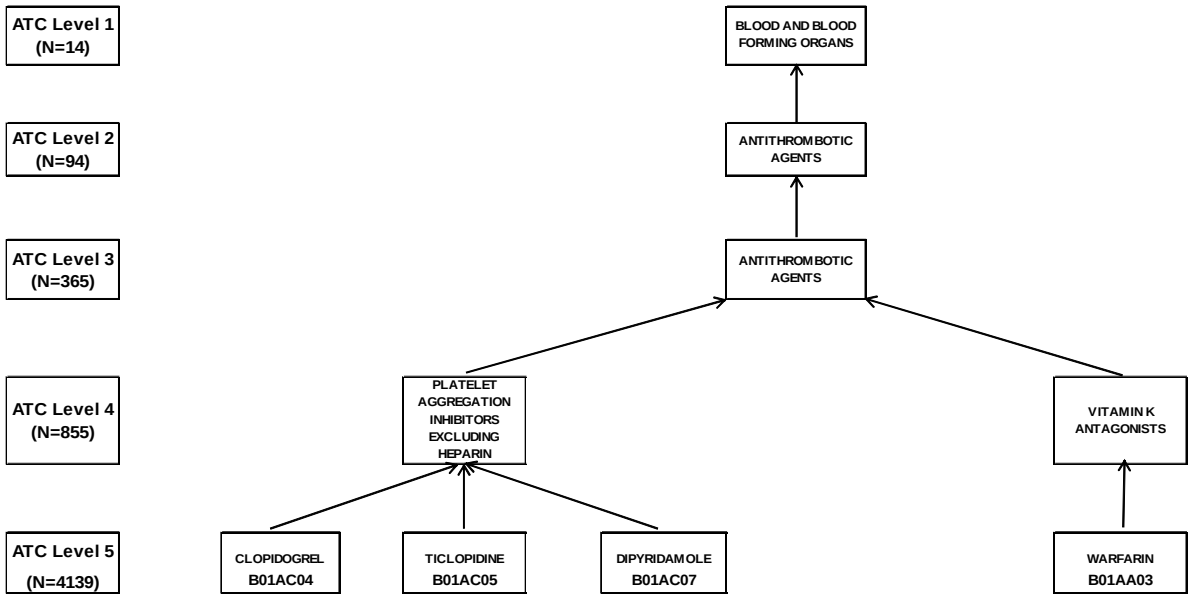

Figure 2.2. An example of aggregations of medications into 5 levels of the Anatomical Therapeutic Chemical (ATC) classification

In general, aggregation of potential covariates into higher-level categories increases the number of covariates that are present in at least 100 observations, the default requirement of the hd-PS, and increases the prevalence of the covariate in exposed and unexposed groups, which increases the covariate's prioritization from the Bross formula if it is associated with treatment [13,26]. But aggregation may simultaneously weaken covariate-exposure and/or covariate-outcome relations, reducing prioritization in the Bross formula. The latter also has the potential to change the impact of control for the aggregated covariate on the adjusted risk ratios. No study to date has assessed how the hd-PS performance is affected by aggregating medical diagnoses and/or medications, especially in cohorts with relatively few patients, rare outcome incidence, or low exposure prevalence.

D. SUMMARY

The hd-PS is a potential algorithm for active drug safety monitoring systems using longitudinal healthcare databases. Prior studies showed that the hd-PS leads to confounding control

that is as least as good as the one obtained by adjustment limited to covariates predefined by expert knowledge [13,15,16]. It could reduce programming time and error, and run in studies pooling multiple claims databases [13,14]. The hd-PS performance has been evaluated with few outcome events or few exposed subjects in small cohorts only [17]. There was a lack of the literature on the potential factors influencing the hd-PS performance.

To our knowledge, this was the first study to evaluate: 1) potential factors affecting the performance of the hd-PS: calendar time periods, data sources, low outcome incidence or exposure prevalence in medium sized or large cohorts; and 2) the effects of aggregation of medical diagnoses into CCS and/or of medications into ATC on the hd-PS adjustment for confounding in cohorts with small size, rare outcome incidence or low exposure prevalence. Our results of this study would add great amount of knowledge to the field and determine the value of the hd-PS.

CHAPTER III

METHODS

This research assesses the factors which can affect the performance of the hd-PS algorithm to adjust confounding for treatments effects using claims databases: different calendar time periods and administrative data sources (S*pecific aim 1*); low outcome incidence or exposure prevalence in medium sized or large cohorts (S*pecific aim 2*); and aggregating medical diagnoses and/or medications, especially in cohorts with relatively few patients, rare outcome incidence, or low exposure prevalence (S*pecific aim 3*). The methods that are common to the three specific aims components will be described, followed by the methods specific to each specific aim. We use a retrospective cohort of upper gastrointestinal (GI) complications with celecoxib versus nonsteroidal anti-inflammatory drugs (NSAIDs) for osteoarthritis (OA) and rheumatoid arthritis (RA) as an example for assessment of the performance of the hd-PS in the cohorts with multiple settings since the upper GI complication treatment effect of COX-2 versus NSAIDs is well established based on several Randomized Controlled Trials (RCT) [18-23]. We therefore assume that a treatment effect estimate closer to 0.5 is less biased by confounding. The current study was exempted by the Institutional Review Board of University of North Carolina at Chapel Hill.

A. DATA SOURCES

We identified incident user cohorts of upper gastrointestinal (UGI) complication with celecoxib versus traditional non-selective non-steroidal anti-inflammatory drugs (tNSAID), ibuprofen or diclofenac, for rheumatoid arthritis (RA) and osteoarthritis (OA) from two large longitudinal administrative databases in the United States (US): MarketScan® commercial claims and encounters of Thomson Reuters Healthcare [65] and Optum® Impact® National Managed Care Benchmark Database [66]. MarketScan is a longitudinal 10-year healthcare claims database which captures

patient demographics, inpatient and outpatient diagnoses and procedures, and medications from a selection of large private employers, health plans, government agencies and other public organizations. Optum is a longitudinal medical claims history for more than 98 million individuals. The Optum data come from more than 46 health plans in the US with available information: patient demographics such as age and gender, diagnoses, procedures, and medications recorded during outpatient visits and hospital admissions.

During the research process, we encountered the issue of incomplete inpatient diagnoses and procedures for years 1999-2002 of the Optum® Impact® National Managed Care Benchmark Database [66]. To compensate for this limitation, we added MarketScan® commercial claims and encounters of Thomson Reuters healthcare [65].

B. METHODS COMMON TO THREE SPECIFIC AIMS

Study population

We extracted data for all patients with an index date (date of first dispensing of celecoxib or a tNSAID) fell between 1 January 2001 – 30 June 2009 (MarketScan) or 1 July 2003 – 30 June 2008 (Optum). These dates were chosen because of availability of data including inpatient diagnoses and procedure. Additional selection criteria were age 18-65, health insurance plan with full medical and pharmacy benefits, at least 6 months of enrollment history at the index date, at least one diagnosis of RA [International Classification of Diseases (ICD-9) code 714, 7140, 7141, 7142, 7143x] or OA (ICD-9 code 715x, 721x); no NSAID dispensing during the 6 months prior to the index date (wash-out period); and no record of any of the following conditions in 6 months prior to the index date: gastrointestinal ulcer disorders, gastrointestinal hemorrhage, active renal, hepatic, coagulation disorders, allergies, malignancy, esophageal or gastroduodenal ulceration.

The study outcome of UGI complication was defined as either first peptic ulcer disease complications including perforation, UGI hemorrhage (ICD-9 code 531x, 532x, 533x, 534x, 535x, 5780), or a physician service code for UGI hemorrhage (Current Procedure Terminology (CPT) code

43255 or ICD-9 procedure code 4443) during a 60-day follow-up period after the initiation of the study drug. These outcome definitions were used in a previous study [13] and validated for 1,762 patients in a hospital discharge database with a positive predictive value of 90% validated against medical chart review [72].

Methods for potential confounder selection

Three different methods were employed to select potential confounders to derive the PS:

(1) *Expert knowledge only.* With this frequently used method, confounders are pre-specified based on the subject-matter knowledge. The *a priori* confounders we selected for this study were age, gender, calendar year, hypertension, congestive heart failure, coronary artery disease, inflammatory bowel disease, prior dispensing of gastroprotective drugs, warfarin, antiplatelet, and oral steroids [13,15,73-75];

(2) *Semi-automated covariate selection*. With this method, we used the hd-PS algorithm to select confounders to supplement those selected based on expert knowledge.

(3) *Automated covariate selection.* This method uses the hd-PS algorithm and a more limited set of *a priori* covariates. We used only age, gender, and calendar year as *a priori* covariates.

Below is a Causal Directed Acyclic Graph (DAG) of celecoxib and upper gastrointestinal complications.

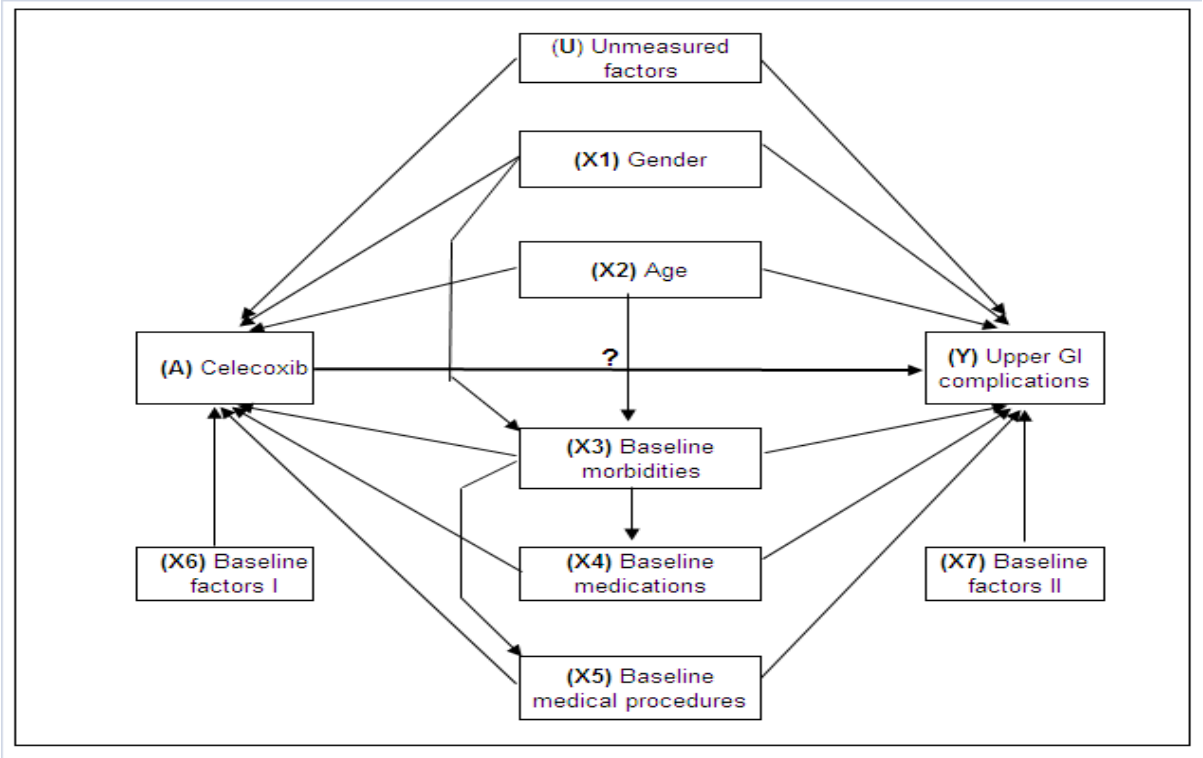

Figure 3.1. A causal Directed Acyclic Graph (DAG) for of celecoxib and Upper Gastrointestinal Complications (→: causes)

Selection of Variables and Parameters for Propensity Score Balancing

The computer algorithm used to implement methods 2 and 3 in our study is the multi-step hd-PS macro [13]. The macro proceeds in these steps: (i) identify data dimensions, e.g., diagnoses, procedures, and medications; (ii) define covariates using the codes within each dimension; (iii) assess candidate covariates by their recurrence, i.e. multiplicity of the same code, (once, sporadic or frequent); (iv) prioritize covariates by their potential for confounding control based on the bivariate associations of each covariate with the treatment and with the outcome according to Bross formula [13,26]; (v) select a pre-specified number of covariates for adjustment; and (vi) estimate of the PS using the selected covariates plus any predefined covariates.

We used hd-PS with the following potential confounders: (i) demographic covariates (sex, continuous age, calendar year) and other baseline (i.e., during six months prior to the index date) pre-specified covariates; (ii) baseline data items from five data dimensions: inpatient diagnoses,

inpatient procedures, outpatient diagnoses, outpatient procedures and outpatient drug use. We allowed a maximum of 500 covariates for the PS adjustment in a multiplicative model and a maximum of 200 variables from each data dimension, based on 5-digit granularity of ICD-9, CPT or Healthcare Common procedure Coding System (HCPCS), and generic drug names.

C. METHOD FOR SPECIFIC AIM 1

For each cohort from each data source, we compared the 60-day risk of UGI complication in celecoxib initiators to that in tNSAID initiators. We used a log-binomial regression model to estimate adjusted risk ratios (RRs) and their 95% confidence intervals (CIs). We estimated RRs instead of odds ratios to avoid the non-collapsibility of the odds ratio under exchangeability [76]. Models included, in addition to the indicator variable for celecoxib initiation the PS in deciles estimated by different sets of covariates identified by the three methods as dummy variables.

We also conducted pooled analyses, a PS-based pooling method [63], for the two combined databases. The pooled analyses were stratified by data source and adjusted by deciles of separate PSs [63]. We employed the PS distribution without trimming of the non-overlap region to create separate PS deciles for each data source cohort. In addition to the use of the propensity score deciles, we applied 1:1 PS greedy matching [64] starting at the eighth decimal place [63] and continued to the first within the data source to create a matched cohort from each data source and then conducted the pooled analyses for these matched cohorts.

D. METHOD FOR SPECIFIC AIM 2

We selected a cohort example of MarketScan, July 2003-September 2004 for resampling to investigate specific aims 2 and 3.

Sampling techniques to generate cohorts with different sizes, outcome incidences and exposure prevalences

The full cohort consisted of 18,829 patients (7,197 prescribed celecoxib and 11,632 prescribed ibuprofen or diclofenac); 117 patients developed a UGI complication. For each

aggregation scenario (including no aggregation), we created six categories of 100 cohorts, as follows. We created "small" cohorts by drawing 50% (category 1) and 20% (category 2) simple random samples, 100 times each, without replacement. We created cohorts with low outcome incidence by drawing 50% (category 3) and 20% (category 4) simple random samples, 100 times each, without replacement, from the 117 cases and re-coding the remaining cases as noncases. Cohorts in categories 3 and 4 consisted of the sampled and recoded cases plus the original 18,712 noncases. Finally, we created cohorts with low exposure prevalences by drawing 50% (category 5) and 20% (category 6) simple random samples, 100 times each, without replacement, from the 7,197 exposed subjects and replacing the unselected exposed subjects with the same number of randomly selected unexposed patients. Cohorts in categories 5 and 6 consisted of the sampled exposed subjects, replacements for the unselected exposed subjects, plus the original 11,632 unexposed subjects.

We applied hd-PS to the full study cohort to estimate the treatment effect and used it as the reference value for comparison with results from the generated cohorts. For the 100 samples in each of the cohort categories, we calculated summary statistics for the estimated risk ratios (geometric mean, 25^{th} and 75^{th} percentiles), the mean percentage of covariates selected by hd-PS in the full cohort that were also selected by hd-PS in the samples, the median number of exposed and unexposed subjects, the median number of exposed and unexposed outcomes.

Simulations to validate sampling techniques

To validate the proposed sampling techniques, we simulated data of 10,000 subjects, 6 covariates independent of one another. We started with 3 binary covariates, X_1, X_2, and X_3, each with a prevalence of 0.2, and 3 continuous covariates, X_4, X_5, and X_6, each with a mean=0 and variance=1. We estimated the predicted probability of the binary intended treatment T with prevalence ~33% based on these 6 covariates and covariate-treatment associations using a logistic model:

$$p(T|X_1\text{-}X_6)=(1+\exp(-(\alpha_0 + \alpha_1X_1 + \alpha_2X_2 + \alpha_3X_3 + \alpha_4X_4 + \alpha_5X_5 + \alpha_6X_6)))^{-1}$$

The number of outcomes Y was assigned from a Poisson distribution based on this expected value.

$$E(Y|T,X_1\text{-}X_6)=\exp(\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6 + \beta_TT))$$

The parameter value for α_0 was selected to obtain a prevalence of T of approximately 33%, the one for β_0 in equation for an incidence of approximately 0.1 per observation over a fixed follow-up time in the untreated, a true exposure RR=0.5. We used parameter values: α_1=0.69; α_2=0; α_3=−1.61; α_4=0.41; α_5=0; α_6=−0.69; β_1=0; β_2=0.69; β_3=−1.61; β_4=0; β_5=0.41; and β_6=−0.69. We used the log-linear outcome model. We simulated 1,000 cohort studies with n =10,000 for each sampling scenario. Results of 1,000 runs at 50%, 20%, 10% and 5% sampling rates demonstrated that our proposed techniques did not affect the treatment effect estimate while being able to keep total sample size constant for outcome incidence or exposure prevalence samplings (Appendix B).

D. METHOD FOR SPECIFIC AIM 3

Aggregations of medical diagnoses and medications

In the basic scenario, we applied the hd-PS with up to 5-digit granularity of ICD-9 for inpatient and outpatient diagnoses. Note that 3-digit ICD-9 codes are kept separate from 4- and 5-digit codes in the hd-PS despite some hierarchy between these levels. We transformed ICD-9 diagnoses into four-level CCS categories via the cross-mapped ICD-9 to CCS multi-level diagnoses table [68]. There are 18, 134, 355 and 207 categories in CCS levels 1, 2, 3 and 4, respectively. However, not all ICD-9 codes have a corresponding CCS code in all four levels. Therefore we created a “universal” CCS by using the most granular code available for each ICD-9 diagnosis code. We separately investigated different levels of ICD-9 granularity by using the first 3- or 4-digit ICD-9 codes.

We aggregated medications to five levels of the Anatomical Therapeutic Chemical (ATC) classification of the World Health Organization (WHO) [69]. This system classifies active substances into different groups based on their target organ or system and their therapeutic, pharmacological and chemical properties. Drugs are classified into fourteen main groups (1st level) with pharmacological or therapeutic subgroups (2nd level). The 3rd and 4th levels are chemical, pharmacological or therapeutic subgroups, and the 5th level is the chemical substance. Several ATC groups are subdivided into both chemical and pharmacological groups. The pharmacological group is often

chosen if a new substance fits in both a chemical and pharmacological 4th level. Substances in the same 4th ATC level are not pharmacotherapeutically equivalent, as they may have different modes of action, therapeutic effects, drug interactions and adverse drug reaction profiles. New 4th levels are commonly established if at least two approved substances fit in the group. A new substance not clearly belonging to any existing group of related substances of ATC 4th level will often be placed in an X group ("other" group).

We evaluated each aggregation scenario by estimating the amount of residual confounding, calculated as the difference in the natural logarithms of the estimated risk ratio and the natural logarithm of 0.50, representing the RCT findings. To estimate the change in residual confounding resulting from each aggregation scenario, we calculated the proportional difference in absolute degree of estimated confounding between that scenario and the basic (no aggregation) scenario. For example, for the 20% exposure prevalence cohorts (category 6), the unadjusted (confounded but otherwise presumptively unbiased) estimate is RRu=0.97, and two confounded (but otherwise presumptively unbiased) estimates are RRc1=0.89 (basic, no aggregation) and RRc2=0.81 (combined diagnostic and medication aggregation). Assuming that the unconfounded (true) value is RRt=0.50, estimated confounding in the basic estimate = |ln(0.89) – ln(0.50)| =0.577; estimated confounding in the combined aggregation estimate = |ln(0.81) – ln(0.50)| =0.482. Thus, the proportional difference in absolute degree of estimated confounding between the two estimates =(0.482-0.577)/0.577=-16.3%. Therefore the combined aggregation estimate is 16.3% less confounded than the basic estimate.

CHAPTER IV

RESULTS: Comparative gastro-intestinal risk of nonsteroidal anti-inflammatory drug classes: a cautionary tale about "automated" pharmacoepidemiology

A. INTRODUCTION

Non-random assignment of therapy in clinical practice can lead to confounding by indication in observational studies of drug effects [77]. Confounding occurs when either clinically assigned or self-selected drug therapies with similar indications are prescribed preferentially to patients with different baseline prognoses. Propensity score methods have been developed as a strategy for controlling confounding in situations with many variables and limited knowledge of how to select from among them.

Propensity score methods are an increasingly used approach to control for measured potential confounders, especially in pharmacoepidemiologic studies of rare outcomes in the presence of many covariates from different data dimensions of administrative healthcare databases [5-7]. Selecting from a very large pool of potential confounders for PS models in healthcare claims data is still a major challenge, however. Methods of selecting variables for PS models based on substantive knowledge have been proposed [4, 8-11,78]. In a nutshell the major challenge is to include a sufficient set of confounders and risk factors for the outcome of interest in the model while avoiding the inclusion of instrumental variables [11,56], variables affected by the treatment or the outcome, and colliders in M-structures [12]. Some software packages for automated variable selection for PS models from a large pool of baseline potential confounders are available [13,57]. The High-Dimensional Propensity Score (hd-PS) algorithm for the automated search of variables for PS models has been developed to improve effect estimates compared with PS models limited to predefined covariates [13]. Improved confounding control using the hd-PS has been reported in selected empirical examples, although the gains were very small [13-16].

Moreover, hd-PS has not developed sophisticated options for users to consider specific subtleties of the data. These might include, but are not restricted to, expected changes in the channelling of drugs at specific calendar time points (e.g., due to publication of landmark trials, black box warnings, 'Dear Doctor' letters, marketing activities by drug makers, new guidelines or policies, etc.) [25,52], differences in study periods, and differences in administrative healthcare databases.

To investigate the effect of different calendar time periods, data sources on the hd-PS performance to control for confounding of treatment effects, we created an empirical example based on prior research [13,79] that observed elevated crude risk ratios likely due to confounding by indication in studies of upper gastrointestinal (UGI) complications in rheumatoid arthritis (RA) or osteoarthritis (OA) patients initiating celecoxib compared to traditional non-steroidal anti-inflammatory agents (tNSAID). Celecoxib has been shown to decrease risk for UGI complications in several randomized clinical trials (RCT) [18-23]. We therefore assume that a treatment effect estimate closer to 0.5 is less biased by confounding. We subdivided the MarketScan and Optum cohorts into two subcohorts according to whether the study drug was initiated before or after 30 September 2004 (hereafter referred to as "withdrawal"), and separate analyses for pre and post withdrawal periods were conducted.

B. METHODS

Data sources

We identified incident user cohorts of upper gastrointestinal (UGI) complication with celecoxib versus tNSAIDs, ibuprofen or diclofenac, for rheumatoid arthritis (RA) and osteoarthritis (OA) from two large longitudinal administrative databases in the United States (US): MarketScan® commercial claims and encounters of Thomson Reuters Healthcare [65] and Optum® Impact® National Managed Care Benchmark Database [66]. MarketScan and Optum capture patient demographics, inpatient diagnoses, outpatient diagnoses, outpatient procedures, and medications recorded during outpatient visits and hospital admissions from a selection of large private employers, health plans, government agencies and other public organizations.

Study population

We extracted data for all patients with an index date (date of first dispensing of celecoxib or a tNSAID) between 1 January 2001 – 30 June 2009 (MarketScan) or 1 July 2003 – 30 June 2008 (Optum). These dates were chosen because of availability of data including inpatient diagnoses and procedures. Additional selection criteria were age 18-65 years, membership in a health insurance plan with full medical and pharmacy benefits, at least 6 months of enrollment history at the index date, at least one diagnosis of RA [International Classification of Diseases (ICD-9) code 714, 7140, 7141, 7142, 7143x] or OA (ICD-9 code 715x, 721x); no NSAID (including aspirin) dispensing during the 6 months prior to the index date (wash-out period); and no record of any of the following conditions in 6 months prior to the index date: gastrointestinal ulcer disorders, gastrointestinal hemorrhage, active renal, hepatic, coagulation disorders, allergies, malignancy, esophageal or gastroduodenal ulceration.

The study outcome of UGI complication was defined as either first peptic ulcer disease complications including perforation, UGI hemorrhage (ICD-9 code 531x, 532x, 533x, 534x, 535x, 5780), or a physician service code for UGI hemorrhage (Current Procedure Terminology [CPT] code 43255 or ICD-9 procedure code 4443) during a 60-day period after the initiation of the study drug. These outcome definitions were used in a previous study [13] and validated in 1,762 patients in a hospital discharge database with a positive predictive value of 90% validated against medical chart review [72].

Methods for confounder selection

Three different methods were employed to select potential confounders to derive the PS: (1) *Expert knowledge only.* With this frequently used method, confounders are pre-specified based on the subject-matter knowledge. The *a priori* confounders we selected for this study were age, gender, calendar year, hypertension, congestive heart failure, coronary artery disease, inflammatory bowel disease, prior dispensing of gastroprotective drugs, warfarin, antiplatelet, and oral steroids based on biological rationale and literature review [13, 73-75]; (2) *Semi-automated covariate selection*. With this

method, we used the hd-PS algorithm to select confounders to supplement those selected based on expert knowledge. (3) *Automated covariate selection.* This method uses the hd-PS algorithm and a more limited set of *a priori* covariates including only age, gender, and calendar year of study drug initiation.

The computer algorithm used to implement methods 2 and 3 in our study is the multi-step hd-PS macro [13]. The macro proceeds in these steps: (i) identify data dimensions, e.g., diagnoses, procedures, and medications; (ii) define covariates using the codes within each dimension; (iii) assess candidate covariates by their recurrence, i.e. multiplicity of the same code, (once, sporadic or frequent); (iv) prioritize covariates by their potential for confounding control based on the bivariate associations of each covariate with the treatment and with the outcome according to Bross formula [13,26]; (v) select a pre-specified number of covariates for adjustment; and (vi) estimate of the PS using the selected covariates plus any predefined covariates.

Statistical analysis

Because of limited data availability and to mimic as closely as possible the intention-to-treat analyses in the trials, we used a prescription reimbursement claim as the treatment measure to compare the 60-day risk of UGI complication in celecoxib initiators to that in tNSAID initiators. We used log-binomial regression models to estimate adjusted risk ratios (RRs) and their 95% confidence intervals (CIs). We estimated RRs instead of odds ratios to avoid the non-collapsibility of the odds ratio under exchangeability [76]. Models included, in addition to the indicator variable for celecoxib initiation the PS in deciles estimated by different sets of covariates identified by the three methods as dummy variables. For 1:1 PS greedy matched cohorts, models included the indicator variable for celecoxib initiation.

We also conducted pooled analyses, a PS-based pooling method [63], for the two combined databases. The pooled analyses were stratified by data source and adjusted by deciles of separate PSs [63]. We employed the PS distribution without trimming of the non-overlap region to create separate PS deciles for each data source cohort. In addition to using the propensity score deciles, we

applied 1:1 PS greedy matching [64] starting at the eighth decimal place [63] and continuing to the first within the data source to create matched cohorts from each data source, and then conducted the pooled analyses for these matched cohorts. The current study was exempt by the Institutional Review Board of University of North Carolina at Chapel Hill.

C. RESULTS

Individual data source analyses

In the MarketScan database, compared to the tNSAID group celecoxib users had more baseline risk factors, particularly the warfarin use (5% vs. 1%) (Table 4.1); had longer durations of study drug use (mean 66 days vs. 31 days), and had a higher incidence of UGI complication (0.7% vs. 0.6%). We observed that all three adjusted estimates were reduced from the crude RR of 1.16 in the direction of the RCT finding, although they remained greater than 1.0. There was a substantial overlap in the PS distribution between the two treatment groups. Unexpectedly, the adjustment using greedy matching with the PS created by the semi-automated covariate selection moved the crude RR away from the RCT finding.

In the Optum database, compared to the tNSAID group celecoxib users also had a greater prevalence of the above risk factors (Table 4.2), longer durations of drug use (mean 53 days vs. 29 days), and a higher incidence of UGI complication within 60 days after drug initiation (0.9% vs. 0.8%). An analogous trend of adjusted estimates from the three covariate selection strategies was observed as with the MarketScan database. The adjustment using greedy matching with the PS based on predefined covariates delivered an estimate very slightly closer to the RCT finding than did greedy matching with the automated and semi-automated covariate selection methods.

Overall pooled analyses

In the overall pooled analysis of both databases, celecoxib initiators were older and had more baseline risk factors for UGI complication than did the tNSAID (diclofenac or ibuprofen) initiators (Table 4.3). Celecoxib initiators had a higher incidence of UGI complication within 60-day following

drug initiation (0.8%) than tNSAID initiators (0.7%) (Table 4.4). We observed similar patterns for adjusted estimates using three variable selection methods. There was substantial overlap in the PS distribution between the two treatment groups. Interestingly, the pooled RR from the greedy matching using the predefined covariate PS delivered an estimate very slightly closer to the RCT finding [18-23] than did the other covariate selection methods (Table 4.4).

Calendar time periods

On 30 September 2004 Merck Inc. announced the voluntary withdrawal of rofecoxib from the US market [24]. We hypothesized that channelling of celecoxib, a drug of the same class as rofecoxib, would be affected by the withdrawal of rofecoxib. Specifically, we hypothesized that the influence of UGI risk on preferential prescribing of coxibs would increase (increased confounding by indication). We therefore subdivided the MarketScan and Optum cohorts into two subcohorts according to whether the study drug was initiated before or after 30 September 2004 (hereafter referred to as “withdrawal”), and separate analyses for pre and post withdrawal periods were conducted.

We identified a more pronounced positive association between warfarin and celecoxib in the post- than in the pre-withdrawal subcohorts in both databases (Tables 4.1-4.3). For the MarketScan database, all adjusted estimates for the pre-withdrawal subcohorts were less than 1.0, but estimates for the post-withdrawal subcohort were not. Greedy matching with the PS from the semi-automated covariate selection produced an adjusted RR (for the post withdrawal subcohort) even greater than the corresponding crude RR. For the Optum database, all adjusted RR were above 1.0 for both subcohorts before and after the withdrawal, except for the RR using the greedy matching with automated covariate selection in the October 2004 - June 2008 subcohort. We created an additional MarketScan subcohort with the study drug initiation between July 2003 and September 2004, the same calendar time periods as for the Optum pre-withdrawal subcohort. In this additional MarketScan subcohort, the semi-automated and automated methods moved the crude RR of 1.05 toward a decreased risk of 0.94 and 0.92, respectively.

Table 4.1: Characteristics of initiators of celecoxib or NSAIDs (ibuprofen or diclofenac) in cohorts 18-65 years old, before and after 1:1 greedy matching based on automated hd-PS covariates, from MarketScan database: age at the date of the first medication use and comorbidities/ use of medications as defined during six months prior to the first study medication use

Characteristics		Before or on 30 September 2004				After 30 September 2004				All			
		Original cohort		After PS matching		Original cohort		After PS matching		Original cohort		After PS matching	
		Celecoxib	Ibuprofen or Diclofenac	Celecoxib	Ibuprofen or Diclofenac	Celecoxib	Ibuprofen or Diclofenac	Celecoxib	Ibuprofen or Diclofenac	Celecoxib	Ibuprofen or Diclofenac	Celecoxib	Ibuprofen or Diclofenac
MarketScan*													
Number of drug initiators	(N)	14,939	19, 917	11,463	11,463	39, 406	103, 308	28,633	28,633	54,345	123,225	40,054	40,054
Age (years)													
Median		56.0	52.0	55.0	55.0	55.0	52.0	55.0	55.0	56.0	52.0	55.0	55.0
Mean		54.4	50.6	53.4	53.3	53.7	50.6	53.2	53.4	53.9	50.6	53.5	53.5
Standard Deviation		8.2	9.7	8.5	8.3	8.4	9.7	8.6	8.4	8.3	9.7	8.5	8.4
18-35	(%)	3.1	8.2	3.7	3.5	3.6	8.3	4.1	4.0	3.5	8.3	3.8	3.7
36-45	(%)	10.9	18.7	13.0	13.0	12.0	18.3	13.2	12.4	11.7	18.3	12.8	12.7
46-55	(%)	33.2	37.0	35.8	37.1	34.9	37.2	35.7	36.2	34.4	37.2	35.3	35.7
56-65	(%)	52.9	36.1	47.5	46.4	49.4	36.2	47.0	47.4	50.4	36.2	48.0	47.8
Female	(%)	60.9	59.0	60.2	59.5	59.2	58.4	59.1	59.0	59.7	58.5	59.5	59.4
Hypertension	(%)	22.6	18.2	20.7	20.4	29.7	25.2	27.0	27.3	27.7	24.1	25.3	25.2
Congestive heart failure	(%)	0.4	0.4	0.3	0.4	0.8	0.6	0.6	0.5	0.7	0.5	0.6	0.5
Coronary artery disease	(%)	3.3	2.5	2.7	2.7	4.3	3.1	3.7	3.7	4.0	3.0	3.4	3.5
Chronic renal disease	(%)	0.5	0.6	0.5	0.6	0.8	0.8	0.8	0.8	0.7	0.7	0.7	0.8
Inflammatory bowel	(%)	0.3	0.2	0.3	0.3	0.7	0.4	0.7	0.4	0.6	0.3	0.6	0.3
Use of gastroprotective	(%)	21.3	19.6	18.5	23.4	22.2	19.0	21.8	24.8	21.9	19.1	21.0	24.8
Use of warfarin	(%)	2.8	1.1	1.7	1.4	6.0	1.1	3.7	2.4	5.1	1.1	3.2	2.1
Use of antiplatelet	(%)	1.7	0.9	1.4	1.2	2.2	1.4	2.1	2.1	2.1	1.3	1.9	1.9
Use of oral steroids	(%)	13.0	11.6	12.6	12.0	15.0	14.6	15.1	16.4	14.5	14.1	14.3	14.8

**:1 January 2001-30 June 2009*

Table 4.2: Characteristics of initiators of celecoxib or NSAIDs (ibuprofen or diclofenac) in cohorts 18-65 years old, before and after 1:1 greedy matching based on automated hd-PS covariates, from Optum database: age at the date of the first medication use and comorbidities/ use of medications as defined during six months prior to the first study medication use

Characteristics		Before or on 30 September 2004				After 30 September 2004				All			
		Original cohort		After PS matching		Original cohort		After PS matching		Original cohort		After PS matchin	
		Celecoxib	Ibuprofen or Diclofenac	Celecoxib	Ibuprofen or Diclofenac	Celecoxib	Ibuprofen or Diclofenac	Celecoxib	Ibuprofen or Diclofenac	Celecoxib	Ibuprofen or Diclofenac	Celecoxib	Ibuprofen or Diclofenac
Optum*													
Number of drug initiators	(N)	8,165	12,257	6,916	6,916	36,083	83, 862	34,554	34,554	44,248	96,119	42,041	42,041
Age (years)													
Median		55.0	51.0	54.0	54.0	54.0	51.0	54.0	54.0	54.0	51.0	54.0	54.0
Mean		53.3	49.7	52.5	52.7	52.3	50.0	52.1	52.2	52.5	50.0	52.2	52.3
Standard Deviation		8.4	9.9	8.6	8.4	8.9	9.7	8.9	8.8	8.8	9.8	8.9	8.8
18-35	(%)	3.6	9.4	4.2	3.9	5.2	8.9	5.4	5.2	4.9	8.9	5.1	5.0
36-45	(%)	14.3	20.8	16.1	14.8	15.5	20.2	16.0	15.7	15.3	20.3	15.9	15.4
46-55	(%)	35.2	37.3	36.6	38.2	36.7	37.8	37.0	37.6	36.5	37.8	36.9	38.0
56-65	(%)	46.9	32.5	43.2	43.1	42.6	33.0	41.7	41.5	43.4	33.0	42.1	41.7
Female	(%)	60.1	57.6	56.8	56.9	57.9	57.4	58.3	58.5	58.3	57.4	57.9	58.3
Hypertension	(%)	30.5	25.2	28.8	29.0	32.6	29.4	32.1	32.2	32.3	28.8	31.6	31.7
Congestive heart failure	(%)	1.1	0.6	0.9	0.8	0.9	0.7	0.9	0.9	1.0	0.7	0.9	0.9
Coronary artery disease	(%)	4.9	3.0	3.8	3.8	4.2	3.5	4.0	4.1	4.4	3.4	4.1	4.0
Chronic renal disease	(%)	1.1	0.8	0.9	0.8	0.9	0.9	0.8	1.0	0.9	0.9	0.9	1.0
Inflammatory bowel disease	(%)	0.8	0.4	0.8	0.5	0.8	0.5	0.8	0.6	0.8	0.5	0.8	0.5
Use of gastroprotective	(%)	19.9	12.3	16.6	16.2	17.1	11.7	16.5	16.5	17.6	11.7	16.7	16.4
Use of warfarin	(%)	3.0	0.7	1.5	1.2	3.5	0.9	2.4	1.8	3.4	0.9	2.2	1.8
Use of antiplatelet	(%)	1.7	0.6	1.1	1.0	1.9	1.1	1.8	1.7	1.9	1.0	1.6	1.6
Use of oral steroids	(%)	11.6	9.4	10.6	10.7	13.5	11.3	13.3	13.5	13.2	11.1	12.9	13.0

**: 1 July 2003-30 June 2008*

Table 4.3: Characteristics of initiators of celecoxib or NSAIDs (ibuprofen or diclofenac) in cohorts 18-65 years old, before and after 1:1 greedy matching based on automated hd-PS covariates, from MarketScan and Optum databases: age at the date of the first medication use and comorbidities/ use of medications as defined during six months prior to the first study medication use

Characteristics		Before or on 30 September 2004				After 30 September 2004				All			
		Original cohort		After PS matching		Original cohort		After PS matching		Original cohort		After PS matching	
		Celecoxib	Ibuprofen or Diclofenac	Celecoxib	Ibuprofen or Diclofenac	Celecoxib	Ibuprofen or Diclofenac	Celecoxib	Ibuprofen or Diclofenac	Celecoxib	Ibuprofen or Diclofenac	Celecoxib	Ibuprofen or Diclofenac
MarketScan* and Optum†													
Number of drug initiators	(N)	23,104	32,174	18,379	18,379	75,489	187,170	63,187	63,187	98,593	219,344	82,095	82,095
Age (years)													
Median		56.0	52.0	55.0	55.0	55.0	52.0	54.0	54.0	55.0	52.0	54.0	54.0
Mean		54.0	50.3	53.1	53.1	53.1	50.3	52.6	52.7	53.3	50.3	52.8	52.9
Standard Deviation		8.3	9.8	8.3	8.3	8.6	9.7	8.8	8.7	8.6	9.7	8.7	8.6
18-35	(%)	3.3	8.7	3.9	3.7	4.4	8.6	4.8	4.7	4.1	8.6	4.5	4.4
36-45	(%)	12.1	19.5	14.2	13.7	13.7	19.1	14.7	14.2	13.3	19.2	14.4	14.1
46-55	(%)	33.9	37.1	36.1	37.5	35.8	37.5	36.4	37.0	35.3	37.4	36.1	36.9
56-65	(%)	50.8	34.7	45.9	45.2	46.2	34.8	44.1	44.2	47.2	34.8	45.0	44.7
Female	(%)	60.6	58.5	58.9	58.5	58.6	57.9	58.7	58.7	59.1	58.0	58.7	58.8
Hypertension	(%)	25.4	20.9	23.7	23.6	31.1	27.1	29.8	30.0	29.8	26.2	28.5	28.5
Congestive heart failure	(%)	0.7	0.5	0.5	0.6	0.9	0.6	0.8	0.7	0.8	0.6	0.8	0.7
Coronary artery disease	(%)	3.9	2.7	3.1	3.1	4.3	3.2	3.9	3.9	4.2	3.2	3.8	3.8
Chronic renal disease	(%)	0.7	0.7	0.7	0.7	0.8	0.8	0.8	0.9	0.8	0.8	0.8	0.9
Inflammatory bowel	(%)	0.5	0.3	0.5	0.4	0.7	0.4	0.8	0.5	0.7	0.4	0.7	0.4
Use of gastroprotective	(%)	20.8	16.8	17.8	20.7	19.8	15.7	18.9	20.3	20.0	15.8	18.8	20.5
Use of warfarin	(%)	2.8	1.0	1.6	1.3	4.8	1.0	3.0	2.1	4.3	1.0	2.7	1.9
Use of antiplatelet	(%)	1.7	0.8	1.3	1.1	2.1	1.3	1.9	1.9	2.0	1.2	1.7	1.7
Use of oral steroids	(%)	12.5	3.6	11.8	11.5	14.3	13.1	14.1	14.8	13.9	11.7	13.6	13.9

**:1 January 2001-30 June 2009; †:1 July 2003-30 June 2008*

Greedy matching using the PS with predefined covariates resulted in the highest percentage of exposed patients who were successfully matched (Table 4.4). Greedy matching using the automated hd-PS covariates did not always produce adjusted RRs closer to the RCT finding, compared with greedy matching using the predefined covariate PS. In some cases for the MarketScan database, greedy matching using PS from either automated or semi-automated covariate selection moved the crude estimates far away from the expected decreased risk.

In the pooled analysis of the pre-withdrawal subcohort, all estimates were smaller than the crude RR (Table 4.4). The automated covariate selection, with either PS deciles or greedy matching, moved the crude RR closest to the RCT finding. In the pooled analysis of the post-withdrawal subcohorts, on the contrary, all adjusted RRs were above 1.0. In all pooled and individual analyses, greedy matching with hd-PS variables plus predefined covariates resulted in estimates further from the RCT result than did greedy matching with only predefined covariates (Table 4.4).

Table 4.4: Risk ratios for the upper gastrointestinal complication within 60 days after the study drug initiation for the cohorts from the two healthcare claims databases MarketScan and Optum by using the three selection strategies for confounders and PS deciles or 1:1 greedy matching

	Celecoxib		Ibuprofen or Diclofenac		Unadjusted	PS deciles			PS matching					
						Predefined	hd-PS*	Predefined + hd-PS	Predefined		hd-PS*		Predefined + hd-PS	
	Initiators N	Outcome N (%)	Initiators N	Outcome N (%)	RR (95% CI)†	RR (95% CI)†	RR (95% CI)†	RR (95% CI)†	RR (95% CI)†	%**	RR (95% CI)†	%**	RR (95% CI)†	%**
MarketScan														
Overall: January 2001-June 2009	54,345	379 (0.70)	123,225	742 (0.60)	1.16 (1.02-1.31)	1.12 (0.99-1.27)	1.07 (0.94-1.22)	1.10 (0.96-1.25)	1.07 (0.92-1.25)	93	1.10 (0.92-1.31)	74	1.24 (1.04-1.48)	73
January 2001-September 2004	14,939	83 (0.56)	19,917	120 (0.60)	0.92 (0.70-1.22)	0.91 (0.68-1.20)	0.87 (0.65-1.17)	0.88 (0.65-1.19)	0.80 (0.59-1.09)	90	0.83 (0.59-1.17)	77	0.86 (0.60-1.21)	75
July 2003-September 2004	7,197	46 (0.64)	11,632	71 (0.61)	1.05 (0.72-1.52)	0.95 (0.65-1.38)	0.92 (0.62-1.37)	0.94 (0.63-1.40)	0.87 (0.57-1.32)	95	0.75 (0.46-1.23)	79	0.93 (0.59-1.56)	77
October 2004-June 2009	39,406	296 (0.75)	103,308	622 (0.60)	1.25 (1.09-1.43)	1.19 (1.03-1.37)	1.13 (0.97-1.30)	1.16 (1.00-1.34)	1.21 (1.02-1.45)	95	1.30 (1.06-1.60)	73	1.29 (1.05-1.59)	72
Optum														
Overall: July 2003-June 2008	44,248	396 (0.89)	96,119	717 (0.75)	1.20 (1.06-1.36)	1.10 (0.97-1.25)	1.09 (0.96-1.24)	1.09 (0.96-1.24)	1.03 (0.89-1.19)	98	1.07 (0.92-1.23)	95	1.10 (0.95-1.28)	95
July 2003-September 2004	8,165	68 (0.83)	12,257	78 (0.64)	1.31 (0.95-1.81)	1.21 (0.87-1.69)	1.20 (0.85-1.70)	1.20 (0.85-1.70)	1.17 (0.81-1.69)	96	1.10 (0.75-1.63)	85	1.27 (0.86-1.87)	85
October 2004-June 2008	36,083	328 (0.91)	83,862	639 (0.76)	1.19 (1.04-1.36)	1.08 (0.95-1.24)	1.07 (0.93-1.23)	1.07 (0.93-1.23)	1.01 (0.86-1.17)	99	0.99 (0.85-1.16)	96	1.02 (0.87-1.20)	96
MarketScan and Optum														
Overall	98,593	775 (0.79)	219,344	1,459 (0.67)	1.18 (1.08-1.29)	1.11 (1.01-1.21)	1.08 (0.98-1.18)	1.09 (0.99-1.19)	1.05 (0.95-1.16)	96	1.08 (0.96-1.21)	83	1.16 (1.03-1.30)	83
January 2001-September 2004	23,104	151 (0.65)	32,174	198 (0.62)	1.06 (0.86-1.31)	1.02 (0.82-1.25)	0.97 (0.76-1.25)	1.00 (0.80-1.26)	0.94 (0.74-1.19)	92	0.94 (0.73-1.22)	80	1.02 (0.79-1.32)	79
July 2003-September 2004	15,362	114 (0.74)	23,889	149 (0.62)	1.18 (0.93-1.52)	1.09 (0.85-1.40)	1.07 (0.82-1.39)	1.09 (0.83-1.40)	1.03 (0.78-1.36)	95	0.95 (0.70-1.29)	82	1.13 (0.83-1.54)	81
September 2004-July 2009	75,489	624 (0.83)	187,170	1,261 (0.67)	1.27 (1.16-1.35)	1.13 (1.03-1.25)	1.10 (0.99-1.21)	1.11 (1.00-1.23)	1.09 (0.97-1.27)	97	1.09 (0.97-1.24)	84	1.11 (0.98-1.26)	84

**hd-PS: high dimensional propensity score; †: Risk Ratio and 95% Confidence Interval; **:Percent of exposed patients matched*

D. DISCUSSION

We examined three different variable selection methods for the control of confounding in analyses with two large healthcare databases: covariate selection based on expert knowledge only, an automated search via the hd-PS algorithm, and the combination of these two approaches. The results of the three methods were similar. In particular, we did not observe a uniform improvement of confounding control with the hd-PS. Analyses taking into account various calendar time periods and data sources led to large differences in estimates.

Individual and pooled data analyses

In the separate and pooled data analyses, using semi-automated or automated methods to select covariates consistently yielded RRs closer to the RCT finding than the crude RR, but all the adjusted RRs were still greater than 1.0. Adjusted RRs above 1.0 were inconsistent with previous results from either separate database analysis [13,15,16] or PS-pooling method [63]. Adding the hd-PS covariates into the predefined covariates produced nearly similar estimates with PS deciles, but worse estimates with PS greedy matching.

Calendar time periods

For both data sources after the withdrawal, the three strategies for selecting potential confounders moved the crude estimate, at least slightly, in the direction of the RCT finding, but all adjusted RRs were greater than 1.0. In the pre-withdrawal subcohorts, all adjusted RRs for the MarketScan database (in contrast to the Optum database) were less than 1.0. This may be explained by little unmeasured confounders prior to the withdrawal followed by stronger, intractable unmeasured confounders after the withdrawal. In other words, prior to the withdrawal, there was little channelling in the MarketScan database beyond the one that we could measure whereas after the withdrawal, the channelling becomes stronger and more difficult to measure. Because of fear of a class effect on the cardiovascular side effects, after the withdrawal coxib was more prescribed to those patients either with more severe inflammation or at highest risk for upper gastrointestinal

complications. Severity of RA/OA and of another risk factors for UGI complication was poorly adjusted in this study.

Data sources

For both data sources after the withdrawal, three strategies to select potential confounders somewhat moved the crude estimate to the direction of RCT finding, but all adjusted RRs were greater than 1.0. In the pre-withdrawal subcohorts, in contrast to the Optum, all adjusted RRs of the MarketScan were less than 1.0. This may be explained by little unmeasured confounders prior to the withdrawal followed by stronger, intractable unmeasured confounders after the withdrawal. In other words, prior to the withdrawal, there was little channelling in the MarketScan database beyond the one that we could measure whereas after the withdrawal, the channelling becomes stronger and more difficult to measure. Because of fear of a class effect on the cardiovascular side effects, after the withdrawal coxib was more prescribed to those patients either with more severe inflammation or at highest risk for upper gastrointestinal complications. Severity of RA/OA and of another risk factors for UGI complication was poorly adjusted in this study. In the MarketScan pre-withdrawal subcohorts, the hd-PS led to confounding control that was as least as good as the one obtained by adjustment limited to covariates predefined by expert knowledge as previously observed [13,15,16]. In these previous studies [13,15,16] the estimated risk reduction of UGI complication after adjustment for investigator-specified covariates and the hd-PS algorithm was 6-21% and 12-22%. Approximately 50% lower risk of UGI complication among celecoxib initiators compared with tNSAID initiators was reported in RCT finding.

In the Optum pre-withdrawal subcohort, on the contrary, the use of three methods to select the potential confounders budged the crude RR of 1.31 to adjusted RRs ~1.20. It is arguable that the three selection strategies still have some benefits to shift the crude estimate into the direction of the RCT finding, nevertheless the adjusted RRs were still above 1.0. It is worth noting that the Optum pre-withdrawal subcohort had the study drug initiation between July 2003 and September 2004, but the MarketScan pre-withdrawal subcohort had the study drug initiation between July 2001 and

September 2004. In the MarketScan subcohort with the study drug initiation between July 2003 and September 2004, both semi-automated and automated selection methods moved the crude RR toward a reduced risk (Table 4.4), regardless of that the latter cohort had 46 UGI events in the celecoxib group.

The higher imbalance of warfarin use in the celecoxib group versus referent group after PS matching suggests the difference in channelling bias for warfarin between post and pre-withdrawal subcohorts (Tables 4.1-4.3). There were potential changes of channelling bias over time in our studies. Different estimates after hd-PS adjustment using the three variable selection methods by calendar time periods and databases were in Figures 4.1 - 4.4.

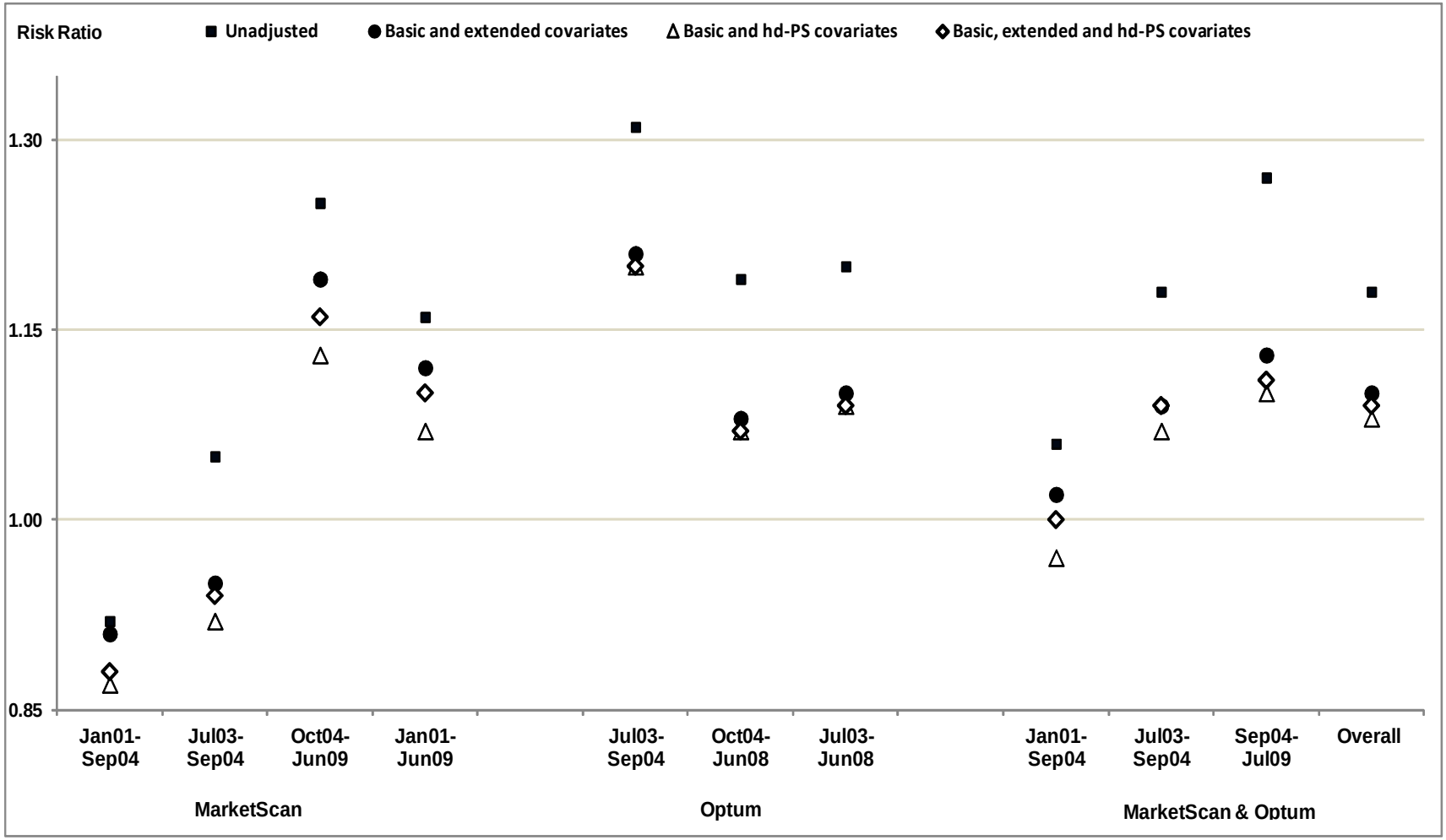

Figure 4.1. Risk Ratios for the upper gastrointestinal complication within 60 days after the study drug initiation for cohorts in the MarketScan and Optum databases by using the hd-PS deciles and three selection strategies for confounders

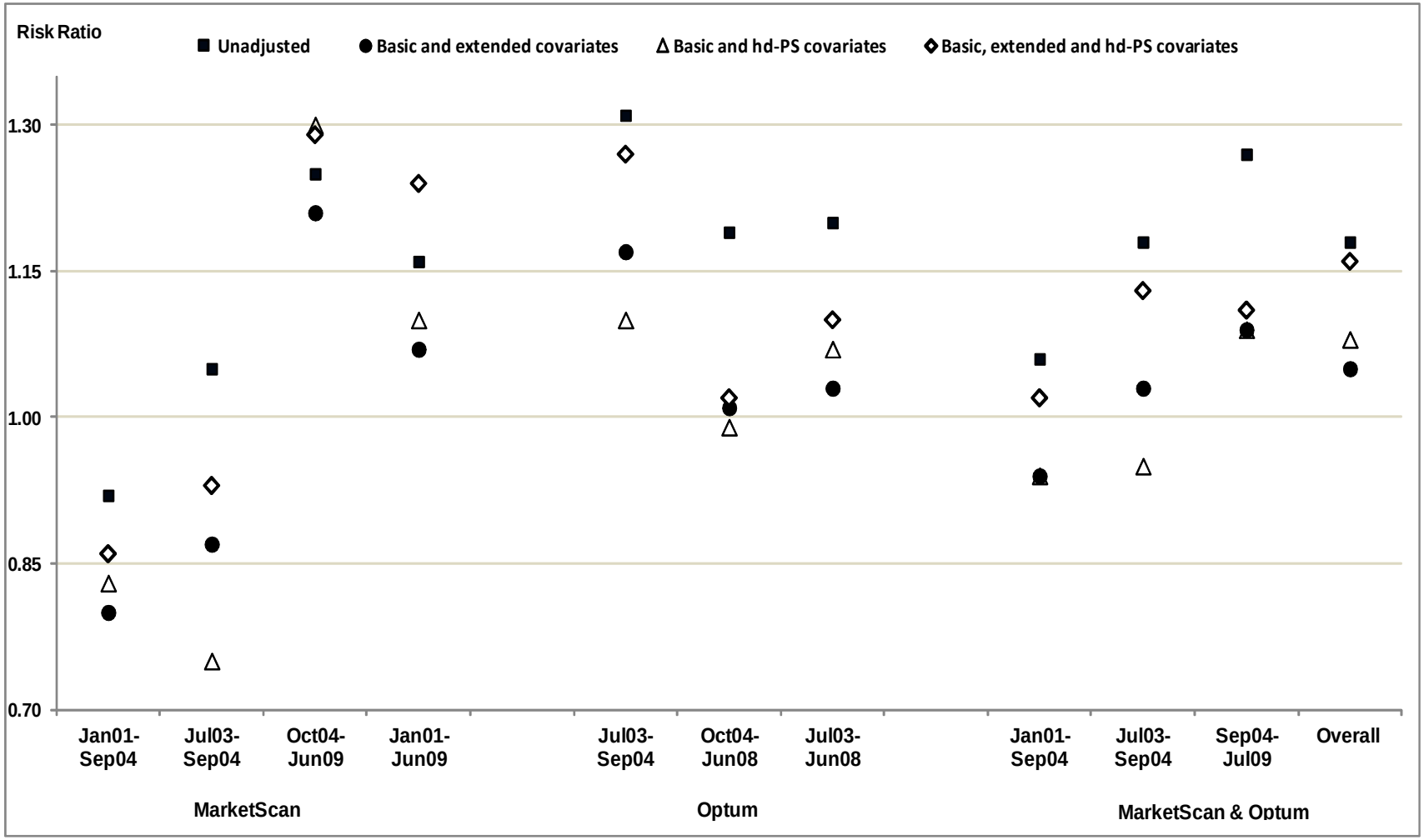

Figure 4.2. Risk Ratios for the upper gastrointestinal complication within 60 days after the study drug initiation for cohorts in the MarketScan and Optum databases by using 1:1 PS greedy matching and three selection strategies for confounders

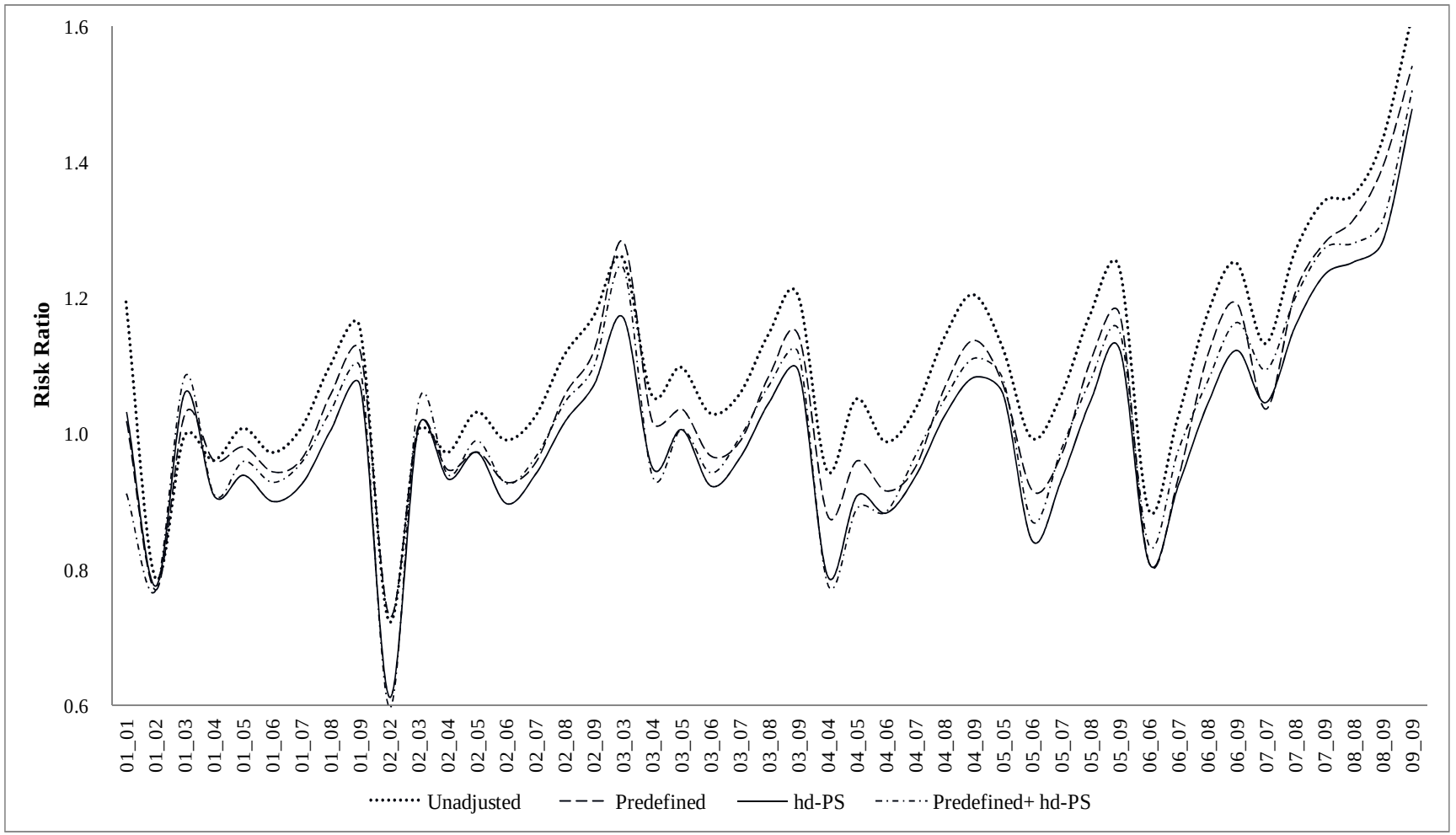

Figure 4.3. Risk Ratios for the upper gastrointestinal complication within 60 days after the study drug initiation for multiple cohorts in the MarketScan database by using the hd-PS deciles and three selection strategies for confounders (*Note: 01-02 means 2001-2002 year cohort)*

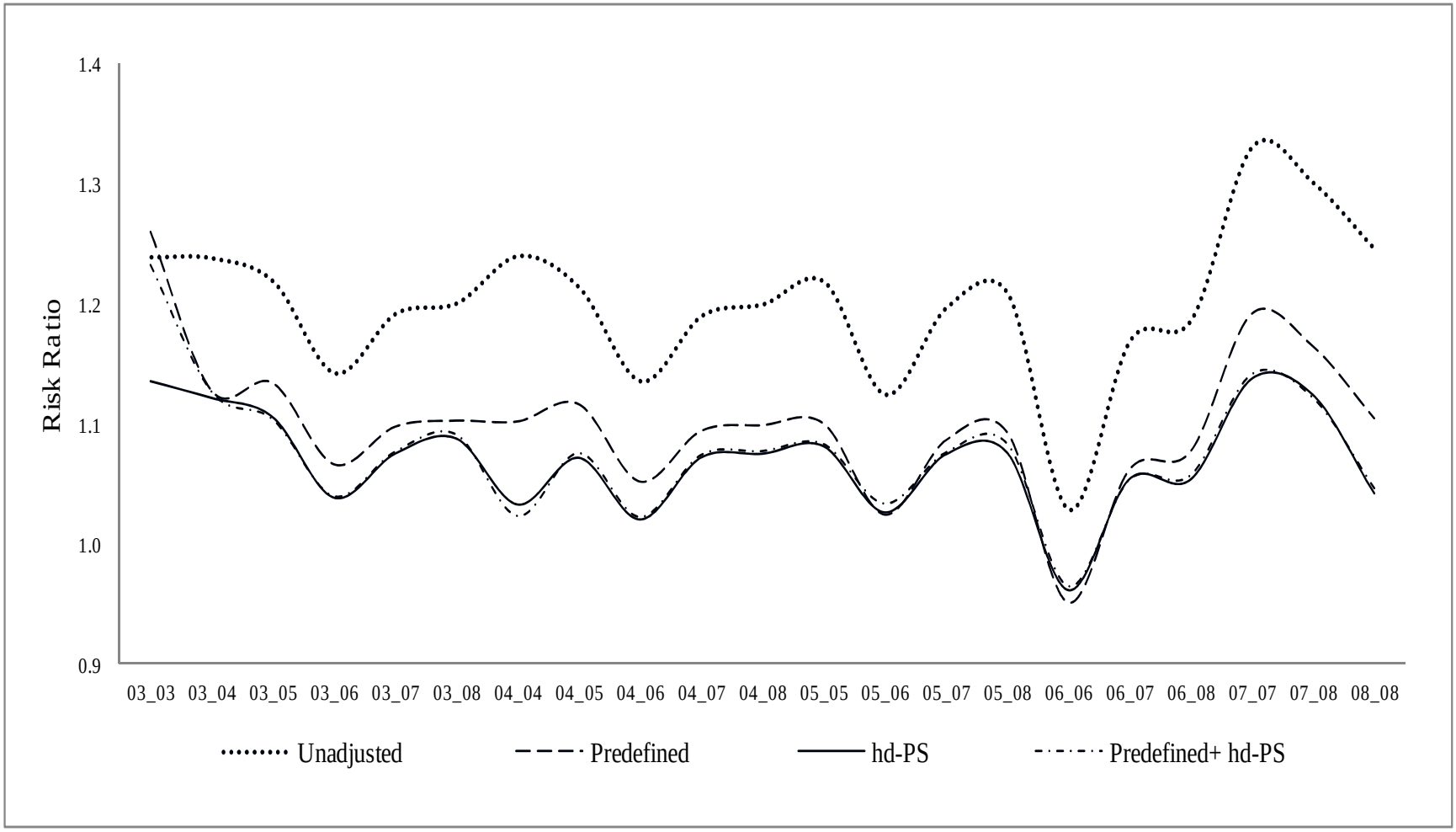

Figure 4.4. Risk Ratios for the upper gastrointestinal complication within 60 days after the study drug initiation for multiple cohorts in the Optum database by using the hd-PS deciles and three selection strategies for confounders (*Note: 03-04 means 2003-2004 year cohort)*

Propensity score deciles and greedy matching

Compared with the estimates using hd-PS deciles, the pooled RR's from PS greedy matching using automated covariates were closer to the RCT finding in the pre-withdrawal subcohorts (Table 4.4). We observed a similar pattern in the MarketScan pre-withdrawal subcohort. Nevertheless greedy matching using PS with semi-automated covariates unexpectedly moved the crude RR away from the RCT finding in MarketScan post-withdrawal subcohort (Table 4.4). In all pooled and individual analyses using greedy matching, adding hd-PS variables to predefined covariates moved the estimate away from the RCT finding. We should cautiously generalize and interpret the results as the retained study populations were different after matching.

To our knowledge, this is the first study to evaluate the impact of various calendar time periods and disparate data sources on treatment effect estimates derived from different methods of selecting confounders. Our evaluation considered various potential influences on hd-PS performance. To explore the possibility of potential overfit of the PS where cohort size or number of outcomes was small, we conducted supplementary analyses in the MarketScan cohorts. First, we tried using hd-PS with only 100 covariates from each data dimension (rather than 200) and a maximum of 200 covariates included in the PS model (rather than 500). In addition, we tried using quintiles for the PS instead of deciles, with both this restricted number of covariates and in the primary analyses. In all instances, results were similar to those from the primary analyses (data not shown).

We also explored other potential influences on treatment effect estimates through channelling bias or other factors, by examining interactions between calendar year and warfarin use, interactions between age and the use of warfarin or antiplatelet drugs, restriction of patients without prior warfarin use, controlling for type of healthcare plan, different numbers of outcomes, truncation of PS distributions. None of these factors could explain the observed difference betweem the crude and adjusted RRs in the pre-withdrawal subcohorts, in either database (data not shown). Results were also similar when we applied the hd-PS with five or three data dimensions for the Optum pre-withdrawal subcohort with drug initiation in July 1999-September 2004, which had missing inpatient

diagnoses and procedures for 1999-2002 (data not shown). We also observed similar results when we included pateients older than 65 years (in the Optum database).

Although the automated covariate selection provided the closest RR in the MarketScan pre-withdrawal subcohort, and a result consistent with a previous study [13] but not with another [15,16], the automated covariate selection should be carefully considered in addition to adjustment for covariates predefined by investigators. Although it is recommended that users of hd-PS should screen and remove instrumental variables and collider bias candidates [11,12, 56], out study focused on the performance of the hd-PS automated covariate selection. Our study can be extended in several ways. Addition of manual review to remove instrumental variables and collider bias candidates might produce improved effect estimates. Research to find ways to automatically identify and remove these kinds of variables would be beneficial to users of the hd-PS. Future studies may also explore alternative approaches to combined analyses of multiple databases by investigating interactions of PS and data source, using Generalized Estimating Equations (GEE) or hierarchical models, or meta-analysis. Using richer information databases such as electronic medical record data to minimize unmeasured confounders should be considered in future studies in order to further investigate the role of the hd-PS algorithm.

Our study has several limitations. First, our study used RCT findings [18-23] as expected treatment effect estimates. We thus empirically compared estimates from different methods and assumed any treatment effect estimates closer to the RCT findings to be less biased by confounding. Our comparison relies on assumptions of no measurement errors in baseline potential confounders. Fully specified simulations with true risk ratios in diversified scenarios could be explored instead of a real-world cohort. Second, it is unclear whether our findings regarding the hd-PS algorithm apply to other treatment-outcome pairs that may be subject to confounding by different factors. Third, studies with few events or small size may have small sample bias or overfit propensity score and outcome models [79,80]. Fourth, the small number of UGI complication cases produced imprecise estimates. Finally, the Optum data did not record inpatient diagnoses and inpatient procedures which occurred

before the year 2003, thus it limited to create a bigger and more comparable cohort with the study drug initiation before 30 September 2004.

In our study, hd-PS added little, if anything, compared with other issues such as channelling bias, differences in calendar time periods, data sources, with which users of “automated” pharmacoepidemiology in active product safety monitoring systems should be cautious of. Different methods of confounder selection by using expert knowledge only, an automated search via the hd-PS algorithm, and both with the propensity score deciles or greedy matching inconsistently reduced confounding by indication to obtain an appropriate effect estimate for studies with various calendar time periods and administrative data sources. The strength of confounding by indication for the effect of non-steroidal anti-inflammatory drugs on upper gastrointestinal complication varied over time before and after the date of voluntary withdrawal of rofecoxib, the same study drug class, from the US market. Users of hd-PS for active product safety monitoring systems should be aware of its benefits and constraints.

CHAPTER V

RESULTS: Effects of Aggregation of Medical Codes on the Performance of the High-Dimensional Propensity Score Algorithm: an Empirical Example

A. INTRODUCTION

Although early detection and assessment of drug safety signals are important [1-3], post-marketing drug safety studies often face challenges such as small size, rare incidence of adverse outcomes, or low exposure prevalence after the launch of a new drug. Nonrandomized studies of treatment effects in healthcare data are vulnerable to confounding bias. Propensity score methods are an increasingly used approach to control for measured potential confounders, especially in pharmacoepidemiologic studies of rare outcomes in the presence of many covariates from different data dimensions of administrative healthcare databases [4-7]. Methods of selecting variables for propensity score models based on substantive knowledge have been proposed [8-12]. However, substantive knowledge may often be lacking, and the meaning of various medical codes may often be unclear [55]. Seeger et al. proposed that health care claims may serve as proxies, in hard-to-predict ways, for important unmeasured variables [14]. Stürmer et al. used propensity score models with over 70 variables representing medical codes present during a baseline period [5]. Johannes et al. created a propensity score model that considered as candidate variables the 100 most frequently occurring diagnoses, procedures, and outpatient medications in healthcare claims [46]. A recently-developed strategy for selecting from a large pool of baseline covariates for propensity score analyses is the use of a computer-applied algorithm [13, 57], such as the High-Dimensional Propensity Score (hd-PS) algorithm. The hd-PS automatically defines and selects variables for inclusion in the propensity score to adjust treatment effect estimates in studies using automated healthcare data [13,15,16].

The hd-PS algorithm prioritizes variables within each data dimension (e.g., inpatient diagnoses, inpatient procedures, outpatient diagnoses, outpatient procedures, dispensed prescription

drugs) by their potential for confounding control based on their prevalence and on bivariate associations of each covariate with the treatment and with the study outcome [13,26]. It excludes variables if those have fewer than 100 patients (exposed and unexposed combined), missing covariate-exposure association or zero/undefined covariate-outcome association. Once variables have been prioritized, a predefined number of variables with the highest potential for confounding per dimension is chosen to be included in the PS.

Combining medical diagnoses or medications into higher-level categories increases the prevalence of the aggregated covariate which may increase the chances of a variable to be selected. In addition to the selection issue, control for a selected aggregated variable may lead to residual confounding. No study to date has assessed how hd-PS performance is affected by aggregating medical diagnoses and/or medications, especially in cohorts with relatively few patients, rare outcome incidence, or low exposure prevalence. To investigate the performance of hd-PS in the cohorts with low outcome incidence or exposure prevalence and the impact of aggregation on hd-PS, we created an empirical example based on prior research [13,78] that observed elevated crude risk ratios likely due to confounding by indication in studies of upper gastrointestinal (UGI) complications in rheumatoid arthritis (RA) or osteoarthritis (OA) patients initiating celecoxib compared to traditional non-steroidal anti-inflammatory agents (tNSAIDs). Celecoxib has been shown to decrease risk for UGI complications in several randomized clinical trials (RCT) by approximately 50% [18-23]. We therefore assume that a treatment effect estimate closer to 0.5 is less biased by confounding.

B. METHODS

Selection of the study cohort

We constructed an incident user cohort [81] to examine UGI complication in RA and OA patients initiating celecoxib or a tNSAID, specifically ibuprofen or diclofenac. All individuals with a first dispensing between 1 July 2003 and 30 September 2004 of celecoxib, ibuprofen, or diclofenac were drawn from the MarketScan® commercial claims and encounters of Thomson Reuters Healthcare [65]. MarketScan is a longitudinal 10-year healthcare claims database which captures patient demographics, inpatient and outpatient diagnoses and procedures, and medications from a selection

of large private employers, health plans, government agencies and other public organizations. We selected patients who as of the date of first dispensing of a study or referent drug (the “index date”) were age 18-65 years, belonged to a health insurance plan with full medical and pharmacy benefits, and had at least 6 months of enrollment history. During the 6 months prior to the index date, patients must have had a diagnosis of RA (ICD-9 code 714, 7140, 7141, 7142, 7143x) or OA (ICD-9 code 715x, 721x) but no NSAID dispensing (including aspirin); and no record of gastrointestinal ulcer disorders, gastrointestinal hemorrhage, active renal, hepatic, coagulation disorders, allergies, malignancy, esophageal or gastroduodenal ulceration.

The study outcome, UGI complication, was defined as either first peptic ulcer disease complications including perforation, an UGI hemorrhage (ICD-9 code 531x, 532x, 533x, 534x, 535x, 5780), or a physician service code for UGI hemorrhage (Current Procedure Terminology (CPT) code 43255 or ICD-9 procedure code 4443). The complication must have occurred during the 60 days after initiation of the study drug. These outcome definitions were validated for 1,762 patients in a hospital discharge database with a positive predictive value of 90% against medical chart review [72].

Aggregations of medical diagnoses and medications

Major U.S. administrative databases represent medical diagnoses with International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9) codes. ICD-9 has its limited hierarchical relationships [67], but the Clinical Classification Software (CCS) developed by the Agency for Healthcare Research and Quality (AHRQ) can be used to aggregate diagnoses into clinically meaningful categories [68]. Similarly, medications, represented by non-hierarchical National Drug Codes (NDC) and generic drug names, can be aggregated using the hierarchical Anatomical Therapeutic Chemical (ATC) drug classification developed by the WHO for drug utilization studies [69].

In the base scenario, we applied the hd-PS with up to 5-digit granularity of ICD-9 for inpatient and outpatient diagnoses. Note that 3-digit ICD-9 codes are kept separate from 4- and 5-digit codes in the hd-PS despite the limited hierarchy between these levels. We transformed ICD-9 diagnoses

into four-level CCS categories via the cross-mapped ICD-9 to CCS multi-level diagnoses table [68]. There are 18, 134, 355 and 207 categories in CCS levels 1, 2, 3 and 4, respectively. However, not all ICD-9 codes have a corresponding CCS code in all four levels. Therefore we created a “universal” CCS by using the most granular code available for each ICD-9 diagnosis code. We separately investigated different levels of ICD-9 granularity by using the first 3- or 4-digit ICD-9 codes.

We aggregated medications to five levels of the Anatomical Therapeutic Chemical (ATC) classification of the World Health Organization (WHO) [69]. This system classifies active substances into different groups based on their target organ or system and their therapeutic, pharmacological and chemical properties. Drugs are classified into fourteen main groups (1st level) with pharmacological or therapeutic subgroups (2nd level). The 3rd and 4th levels are chemical, pharmacological or therapeutic subgroups, and the 5th level is the chemical substance. Several ATC groups are subdivided into both chemical and pharmacological groups. The pharmacological group is often chosen if a new substance fits in both a chemical and pharmacological 4th level. Substances in the same 4th ATC level are not pharmacotherapeutically equivalent, as they may have different modes of action, therapeutic effects, drug interactions and adverse drug reaction profiles. New 4th levels are commonly established if at least two approved substances fit in the group. A new substance not clearly belonging to any existing group of related substances of ATC 4th level will often be placed in an X group ("other" group).

Sampling techniques to generate cohorts with different sizes, outcome incidences and exposure prevalences

The full cohort consisted of 18,829 patients (7,197 prescribed celecoxib and 11,632 prescribed ibuprofen or diclofenac); 117 patients developed an UGI complication. For each aggregation scenario (including no aggregation), we created six categories of 100 cohorts, as follows. We created “small” cohorts by drawing 50% (category 1) and 20% (category 2) simple random samples, 100 times each, without replacement. We created cohorts with low outcome incidence by drawing 50% (category 3) and 20% (category 4) simple random samples, 100 times each, without replacement, from the 117 cases and re-coding the remaining cases as noncases. Cohorts in

categories 3 and 4 consisted of the sampled and recoded cases plus the original 18,712 noncases. Finally, we created cohorts with low exposure prevalences by drawing 50% (category 5) and 20% (category 6) simple random samples, 100 times each, without replacement, from the 7,197 exposed subjects and replacing the unselected exposed subjects with the same number of randomly selected unexposed patients. Cohorts in categories 5 and 6 consisted of the sampled exposed subjects, replacements for the unselected exposed subjects, plus the original 11,632 unexposed subjects.

The hd-PS algorithm

We implemented the hd-PS algorithm with five data dimensions commonly available in automated healthcare databases: pharmacy claims, outpatient diagnoses, outpatient procedures, inpatient diagnoses, and inpatient procedures. The algorithm identifies the top n most prevalent variables within each data dimension by creating binary variables for each diagnosis, procedure and medication. The prevalence of each variable depends on the granularity of the coding. Each variable is assessed for 3 levels of its within-patient frequency of occurrence: once, sporadic ≥ median number of times, or frequent ≥ 75^{th} percentile number of times [13]. With the default setting of 200 variables for each dimension, 3,000 indicator variables (200 x 3 levels x 5 dimensions) are then prioritized according to their potential for confounding control based on their prevalence and their bivariate associations with the treatment and with the study outcome, according to the Bross formula [13,26]. By default, the top k=500 indicator variables are selected for the propensity score.

Statistical analysis

The hd-PS algorithm can combine these automatically selected variables with predefined covariates chosen by the investigator. Besides a crude model, with no covariates, we fit four log binomial models that adjusted for (1) basic covariates (age [continuous], gender, calendar year of drug initiation), (2) basic plus extended variables (hypertension, congestive heart failure, coronary artery disease, inflammatory bowel disease, prior dispensing of gastroprotective drugs, warfarin, antiplatelet, and oral steroids), selected based on biological rationale and use in the literature [13, 15, 73-75,78], (3) basic plus variables automatically selected with hdPS, and (4) basic, extended, and

automatically selected variables. In the base scenario, we used up to 5-digit granularity of ICD-9, CPT or Healthcare Common Procedure Coding System (HCPCS), and generic drugs. We then re-fitted all models in eight scenarios for aggregation of diagnoses, six scenarios for aggregation of medications, and one scenario that combined the diagnosis and medication aggregations that appeared to perform best across the six categories of cohort samples.

We applied hd-PS to the full study cohort to estimate the treatment effect and used it as the reference value for comparison with results from the generated cohorts. For the 100 samples in each of the cohort categories, we calculated summary statistics for the estimated risk ratios (geometric mean, 25^{th} and 75^{th} percentiles): the mean percentage of covariates selected by hd-PS in the full cohort that were also selected by hd-PS in the samples; the median number of exposed and unexposed subjects; the median number of exposed and unexposed outcomes. We evaluated each aggregation scenario by estimating the amount of residual confounding, calculated as the difference in the natural logarithms of the estimated risk ratio and the natural logarithm of 0.50, representing the RCT findings. To estimate the change in residual confounding resulting from each aggregation scenario, we calculated the proportional difference in absolute degree of estimated confounding between that scenario and the base (no aggregation) scenario. For example, for the 20% exposure prevalence cohorts (category 6), the unadjusted (confounded but otherwise presumptively unbiased) estimate is RRu=0.97, and two confounded (but otherwise presumptively unbiased) estimates are RRc1=0.89 (base, no aggregation) and RRc2=0.81 (combined diagnostic and medication aggregation). Assuming that the unconfounded (true) value is RRt=0.50, estimated confounding in the base estimate = $|\ln(0.89) - \ln(0.50)| = 0.577$; estimated confounding in the combined aggregation estimate = $|\ln(0.81) - \ln(0.50)| = 0.482$. Thus, the proportional difference in absolute degree of estimated confounding between the two estimates $=(0.482-0.577)/0.577=-16.3\%$. We would conclude that the combined aggregation estimate is 16.3% less confounded than the base estimate.

Because of limited data availability, and to mimic as closely as possible the intention-to-treat analyses in the trials, we used a prescription reimbursement claim as the treatment measure. The

current study was exempt by the Institutional Review Board of University of North Carolina at Chapel Hill.

C. RESULTS

In the full cohort, there were 7,197 (38%) celecoxib and 11,632 (62%) ibuprofen or diclofenac initiators with 46 and 71 UGI events, respectively. Celecoxib users were older and had more risk factors for UGI complications than did the tNSAIDs users (Table 5.1). The RR for UGI complication associated with celecoxib versus tNSAIDs was 1.05 in the crude model, compared to 0.92 in the model that used hd-PS automated variable selection (along with the basic covariates) (Table 5.2). Consistent with the sampling procedures described above, the median numbers of patients in cohorts in categories 1 and 2 were about 3,594 and 1,441, respectively, the median outcome incidence proportions in categories 3 and 4 were about 0.32% and 0.14%, respectively, and the median exposure prevalences in categories 5 and 6 were about 19% and 8%, respectively.

In all cohort categories except category 2, where the total study size was only about 3,790, the geometric means of the hd-PS adjusted risk ratios were similar to the full cohort risk ratios. This similarity held even in cohort categories 4 and 6, where the number of exposed patients with an outcome event was approximately 10. In all categories except category 6, where the exposure prevalence was only 8%, the geometric means of the hd-PS adjusted risk ratios were at least slightly closer to the RCT finding than the geometric means of the risk ratios adjusted for only the basic and extended covariates. A majority of the covariates that hd-PS identified in the full cohort were also selected by hd-PS in the samples in categories 1,3, and 5, where the number of exposed outcomes was at least 20, but also in category 6, where there were only 10 exposed outcomes but a large total number of outcomes.

A scenario with combined aggregations of medications into ATC level 4 and of diagnoses into CCS level 1 consistently performed best, reducing residual confounding from 8.9% to 19.3% compared to the base scenario (Table 5.3). When we experimented with different aggregations for

diagnoses, without any aggregation for medications, aggregating ICD-9 diagnosis codes into different CCS levels inconsistently changed the adjusted risk ratios. Note that in our empirical setting not controlling for any measure of co-morbidity resulted in the estimate closest to the RCT finding (Table 5.3). When we aggregated ICD-9 diagnosis codes into CCS levels 1 or 2, the adjusted risk ratios in the samples were generally closer to the RCT finding. In contrast, aggregations of ICD-9 codes into CCS universal, CCS level 3, CCS level 4, or 3- or 4-digit ICD-9 categories did not improve the adjusted point estimates.

Aggregating medications into chemical, pharmacological or therapeutic subgroups of ATC level 4, slightly improved adjusted estimates in all cohort categories except category 4, the 20% outcome incidence samples. In contrast, aggregations of medications into categories of the other ATC levels produced nearly the same or even worse adjusted risk ratios in all cohort categories.

Table 5.1. Characteristics of Initiators of Celecoxib or NSAIDs (ibuprofen or diclofenac) in a Cohort 18-65 Years Old Between 1 July 2003 and 30 September 2004 of MarketScan Database: Age at the Date of the First Medication Use and Comorbidities/ Use of Medications as Defined During Six Months Prior to the First Medication Use

Characteristics	Celecoxib N=7,197 (38%)		Ibuprofen or Diclofenac N=11,632 (62%)	
	N	%	N	%
Age (years)				
Median	56.0		52.0	
Mean	54.1		50.4	
Standard Deviation	8.2		9.7	
18-35	235	3.3	996	8.6
36-45	854	11.9	2,164	18.6
46-55	2,373	33.0	4,339	37.3
56-65	3,735	51.9	4,133	35.5
Female	4,387	61.0	6,869	59.1
Hypertension	1,748	24.3	2,191	18.8
Congestive heart failure	36	0.5	56	0.5
Coronary artery disease	270	3.8	297	2.6
Chronic renal disease	44	0.6	59	0.5
Inflammatory bowel disease	26	0.4	30	0.3
Use of gastroprotective drugs	1,567	21.8	2,111	18.1
Use of warfarin	220	3.1	128	1.1
Use of antiplatelet	143	2.0	108	0.9
Use of oral steroids	963	13.4	1,356	11.7

Table 5.2. Geometric Mean of Risk Ratios and a Summary Analysis for Different Cohort Size, Outcome Incidence and Exposure Prevalence of Initiators of Celecoxib or NSAIDs (ibuprofen or diclofenac) in a Cohort 18-65 Years Old Between 1 July 2003 and 30 September 2004 of MarketScan Database

Cohort and Variable Selection Method	Median of exposed subjects N (%)	Median of exposed outcomes N (%)	Median of unexposed subjects N (%)	Median of unexposed outcomes N (%)	Geometric mean of RR (a)	25th-75th percentiles of RR of samples (b)	Mean variable coverage % (c)
Full Cohort[d]	7197 (38)	46 (0.64)	11632 (62)	71 (0.61)			
Unadjusted					1.05		
Basic covariates					0.98		
Basic and extended covariates					0.95		
Basic and hd-PS covariates					0.92		100
Basic, extended and hd-PS covariates					0.94		100
Category 1: 50% Size Sample	3594 (38)	23 (0.64)	5821 (62)	36 (0.62)			
Unadjusted					1.02	0.89, 1.20	
Basic covariates					0.96	0.84, 1.11	
Basic and extended covariates					0.92	0.80, 1.09	
Basic and hd-PS covariates					0.88	0.74, 1.07	65
Basic, extended and hd-PS covariates					0.89	0.74, 1.11	65
Category 2: 20% Size Sample	1441 (38)	10 (0.66)	2325 (62)	14 (0.60)			
Unadjusted					1.10	0.89, 1.37	
Basic covariates					1.03	0.82, 1.29	
Basic and extended covariates					0.99	0.79, 1.24	
Basic and hd-PS covariates					0.94	0.71, 1.21	41
Basic, extended and hd-PS covariates					0.95	0.70, 1.25	41
Category 3: 50% Outcome Incidence Sample	7220 (38)	23 (0.32)	11667 (62)	36 (0.31)			
Unadjusted					1.02	0.89, 1.19	
Basic covariates					0.96	0.84, 1.13	
Basic and extended covariates					0.93	0.81, 1.09	
Basic and hd-PS covariates					0.90	0.78, 1.08	65
Basic, extended and hd-PS covariates					0.91	0.78, 1.08	65
Category 4: 20% Outcome Incidence Sample	7233 (38)	10 (0.14)	11689 (62)	14 (0.12)			
Unadjusted					1.00	0.81, 1.37	
Basic covariates					0.94	0.73, 1.25	
Basic and extended covariates					0.91	0.69, 1.19	
Basic and hd-PS covariates					0.85	0.69, 1.17	42
Basic, extended and hd-PS covariates					0.86	0.70, 1.14	42
Category 5: 50% Exposure Prevalence Sample	3599 (19)	22 (0.61)	15230 (81)	95 (0.62)			
Unadjusted					1.02	0.93, 1.13	
Basic covariates					0.94	0.86, 1.05	
Basic and extended covariates					0.91	0.83, 1.02	
Basic and hd-PS covariates					0.88	0.79, 0.98	81
Basic, extended and hd-PS covariates					0.88	0.79, 1.00	81
Category 6: 20% Exposure Prevalence Sample	1440 (8)	9 (0.63)	17389 (96)	108 (0.62)			
Unadjusted					0.97	0.77, 1.24	
Basic covariates					0.89	0.72, 1.15	
Basic and extended covariates					0.86	0.70, 1.08	
Basic and hd-PS covariates					0.89	0.73, 1.13	73
Basic, extended and hd-PS covariates					0.89	0.72, 1.14	73

Abbreviations: basic covariates included continuous age, gender and calendar year; hd-PS, high-dimensional propenisty score; extended, clinically pre-specified covariates; hd-PS covariates, variables automatically selected by hd-PS; RR, risk ratio.

[a] Geometric mean of the risk ratio observed in 100 samples at this sampling rate.

[b] 25th and 75th percentiles of the risk ratios observed in 100 samples at this sampling rate.

[c] Mean percentage of hd-PS variables in the full cohort also identified in samples.

[d] For the full cohort, all values are the numbers, not mean.

Table 5.3. Geometric Mean of Risk Ratios for Different Cohort Size, Outcome Incidence and Exposure Prevalence of Initiators of Celecoxib or NSAIDs (ibuprofen or diclofenac) in a Cohort 18-65 Years Old Between 1 July 2003 and 30 September 2004 of MarketScan Database by Using the High-Dimensional Propensity Score (hd-PS) Adjustment with Different Aggregation Methods

Cohort and Variable Selection Method		Medical Diagnoses								Medications						Combined
			CCS[a] Level					ICD-9[b]			ATC[c] Level					CCS 1st +
	Base	No Dx	1st	2nd	3rd	4th	Universal	3-digit	4-digit	No Rx	1st	2nd	3rd	4th	5th	ATC 4th
Full Cohort																
Unadjusted	1.05															
Basic covariates	0.98															
Basic and extended covariates	0.95															
Basic and hd-PS covariates	0.92	0.88	0.90	0.89	0.92	0.92	0.94	0.95	0.94	0.94	0.93	0.92	0.92	0.90	0.91	0.85
%[d]		-7.0	-3.7	-4.4	0.1	1.0	3.6	5.1	4.1	3.9	2.6	0.0	0.8	-2.9	-1.4	-12.1
Basic, extended and hd-PS covariates	0.94	0.91	0.91	0.92	0.95	0.94	0.96	0.96	0.95	0.91	0.96	0.94	0.94	0.90	0.93	0.88
%[d]		-5.0	-4.4	-2.5	1.0	0.6	3.6	4.0	2.1	-5.0	3.7	-0.5	-0.7	-6.0	-1.3	-10.9
Indicator Variables (k=500)																
Outpatient Diagnoses (N)	136	0	32	90	97	54	123	133	139	224	198	177	154	144	133	34
Inpatient Diagnoses (N)	9	0	22	18	19	5	16	14	11	12	11	11	9	9	7	23
Medications (N)	167	247	216	186	181	213	171	166	163	0	36	76	122	148	177	194
Outpatient Procedures (N)	152	210	188	166	163	187	153	151	151	220	211	194	174	161	148	206
Inpatient Procedures (N)	36	43	42	40	40	41	37	36	36	44	44	42	41	38	35	43
Category 1: 50% Size Sample																
Unadjusted	1.02															
Basic covariates	0.96															
Basic and extended covariates	0.92															
Basic and hd-PS covariates	0.88	0.85	0.83	0.85	0.88	0.85	0.89	0.88	0.91	0.90	0.89	0.88	0.88	0.87	0.88	0.83
%[d]		-5.4	-9.2	-5.8	0.0	-5.5	2.2	1.2	6.7	4.8	3.0	1.5	1.0	-1.5	1.7	-9.9
Basic, extended and hd-PS covariates	0.89	0.87	0.85	0.86	0.89	0.87	0.90	0.90	0.92	0.89	0.90	0.89	0.89	0.88	0.90	0.84
%[d]		-3.5	-7.4	-5.0	0.0	-3.4	2.4	1.5	5.8	0.5	3.0	1.0	1.5	-0.9	1.8	-8.9
Category 2: 20% Size Sample																
Unadjusted	1.10															
Basic covariates	1.03															
Basic and extended covariates	0.99															
Basic and hd-PS covariates	0.94	0.92	0.89	0.90	0.94	0.92	0.96	0.94	0.99	0.97	0.98	0.96	0.95	0.93	0.95	0.87
%[d]		-3.8	-9.1	-7.5	0.0	-3.0	2.3	-0.6	8.0	4.5	5.7	2.3	1.5	-1.4	0.8	-12.0
Basic, extended and hd-PS covariates	0.95	0.94	0.89	0.91	0.95	0.94	0.96	0.95	1.00	0.98	0.99	0.96	0.96	0.94	0.95	0.88
%[d]		-1.4	-9.5	-6.3	0.0	-1.4	1.8	-0.3	7.5	5.1	6.6	1.3	1.6	-1.8	-0.1	-11.9

Category 3: 50% Outcome Incidence Sample																
Unadjusted	1.02															
Basic covariates	0.96															
Basic and extended covariates	0.93															
Basic and hd-PS covariates	0.90	0.85	0.84	0.87	0.90	0.86	0.91	0.92	0.85	0.92	0.90	0.89	0.89	0.88	0.91	0.84
%[d]		-8.8	-10.5	-5.8	0.0	-7.6	3.6	4.5	-8.1	3.9	1.3	-0.9	-1.0	-2.4	2.0	-11.9
Basic, extended and hd-PS covariates	0.91	0.87	0.85	0.87	0.91	0.87	0.92	0.92	0.86	0.90	0.91	0.89	0.90	0.89	0.91	0.85
%[d]		-6.5	-10.2	-6.3	0.0	-6.1	3.2	3.4	-8.9	-1.6	0.8	-2.2	-1.3	-3.2	0.5	-11.3
Category 4: 20% Outcome Incidence Sample																
Unadjusted	1.00															
Basic covariates	0.94															
Basic and extended covariates	0.91															
Basic and hd-PS covariates	0.85	0.85	0.82	0.84	0.85	0.85	0.89	0.88	0.85	0.87	0.88	0.86	0.86	0.86	0.87	0.81
%[d]		-1.5	-7.8	-3.2	0.0	-0.7	6.7	6.3	0.0	3.6	4.8	1.1	1.1	0.9	2.9	-10.4
Basic, extended and hd-PS covariates	0.86	0.86	0.83	0.84	0.86	0.87	0.89	0.89	0.86	0.87	0.89	0.87	0.87	0.87	0.86	0.82
%[d]		0.0	-7.5	-4.2	0.0	1.4	6.5	5.5	0.0	2.1	5.8	2.5	2.0	2.0	1.0	-9.8
Category 5: 50% Exposure Prevalence Sample																
Unadjusted	1.02															
Basic covariates	0.94															
Basic and extended covariates	0.91															
Basic and hd-PS covariates	0.88	0.86	0.86	0.88	0.90	0.88	0.90	0.89	0.90	0.90	0.90	0.89	0.87	0.84	0.88	0.81
%[d]		-5.5	-5.0	-0.7	3.3	-0.8	4.2	2.0	4.3	2.9	3.9	1.4	-2.1	-8.0	-1.5	-14.4
Basic, extended and hd-PS covariates	0.88	0.87	0.86	0.88	0.90	0.89	0.91	0.89	0.90	0.89	0.91	0.90	0.89	0.86	0.88	0.82
%[d]		-2.5	-4.1	-0.3	4.2	0.7	5.5	1.8	3.7	2.3	6.3	3.4	0.7	-4.8	-0.6	-12.7
Category 6: 20% Exposure Prevalence Sample																
Unadjusted	0.97															
Basic covariates	0.89															
Basic and extended covariates	0.86															
Basic and hd-PS covariates	0.89	0.83	0.83	0.87	0.88	0.85	0.88	0.88	0.89	0.87	0.87	0.87	0.85	0.83	0.88	0.79
%[d]		-10.8	-10.5	-4.0	-1.2	-6.5	-0.4	-1.6	1.6	-3.1	-3.3	-2.8	-8.0	-10.5	-1.9	-19.3
Basic, extended and hd-PS covariates	0.89	0.84	0.84	0.86	0.88	0.86	0.89	0.88	0.89	0.87	0.88	0.87	0.85	0.85	0.88	0.81
%[d]		-9.8	-10.0	-4.8	-1.3	-5.6	0.6	-1.4	1.3	-2.8	-1.4	-2.9	-6.5	-7.0	-1.6	-16.3

Abbreviations: basic covariates included continuous age, gender and calendar year; extended, clinically pre-specified covariates ; hd-PS, high-dimensional propenisty score; hd-PS covariates, variables automatically selected by the hd-PS algorithm.

Base: the scenario using up to 5-digit ICD-9, procedures, generic drugs for five data dimensions of the hd-PS.
No Dx: the scenario using procedures and generic drugs for three data dimensions of the hd-PS.
No Rx: the scenario using 5-digit ICD-9 and procedures for 4 data dimensions of the hd-PS.

[a] CCS: Four levels of the Clinical Classification Software; Universal, the most granular CCS code available for each ICD-9 code.

[b] ICD-9: International Classification of Diseases, 9th Revision, Clinical Modification.

[c] ATC: 5 levels of the Anatomical Therapeutic Chemical classification.

[d] % change of the residual confounding of aggregation method versus the base scenario at the same variable selection method on the natural log scale with RCT finding of 0.5. The presumptive amount of confounding in the base scenario A = |ln(adjusted RR) – ln(RCT finding)|; in each aggregation method B = |ln(adjusted RR) – ln(RCT finding)|; and C = (B – A)/A

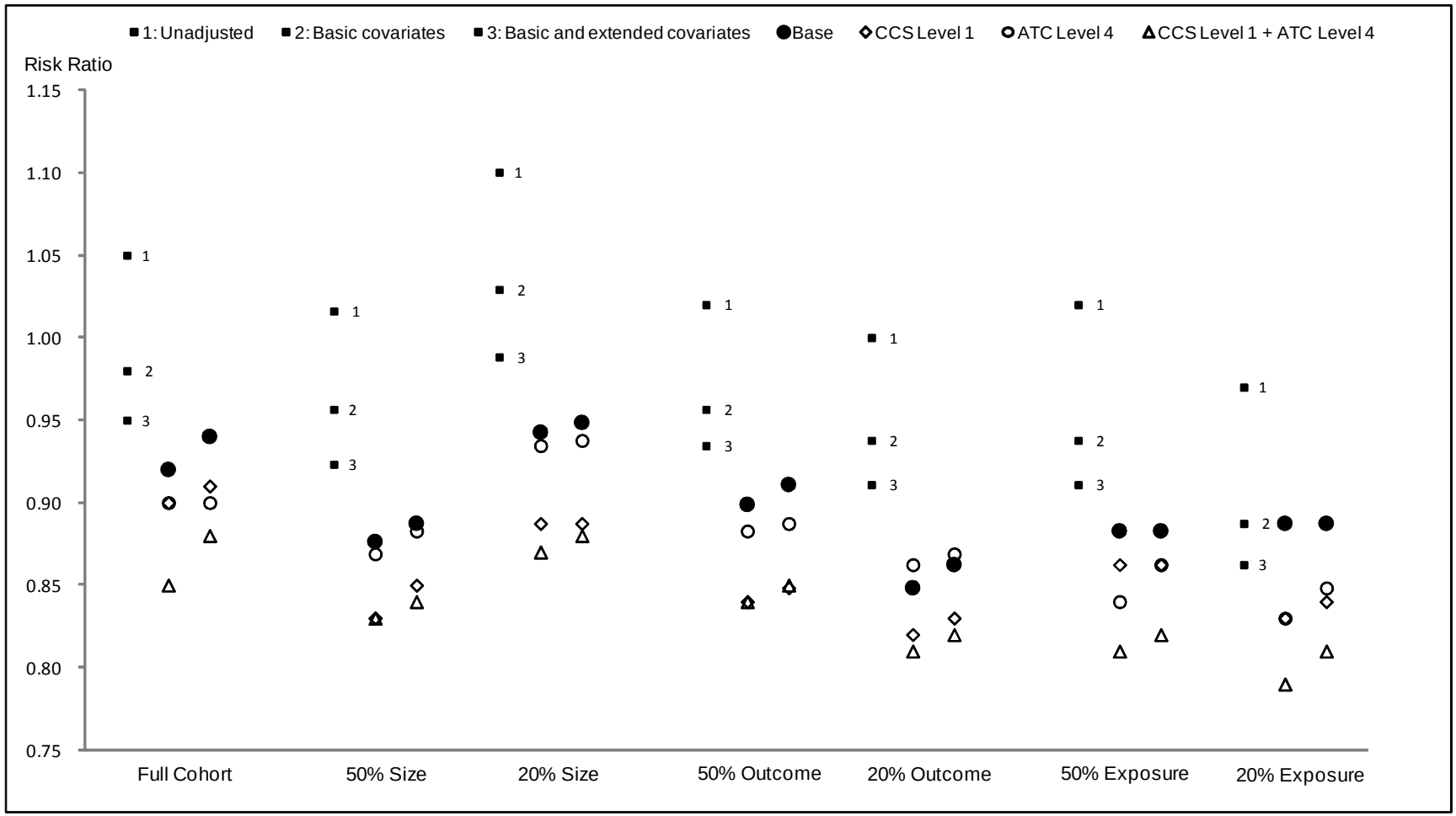

Figure 5.1. Geometric Mean of Risk Ratios for Different Cohort Size, Outcome Incidence and Exposure Prevalence of Initiators of Celecoxib or NSAIDs (ibuprofen or diclofenac) in Cohorts 18-65 Years Old Between 1 July 2003 and 30 September 2004 Using the High-Dimensional Propensity Score Adjustment for Basic and hd-PS Covariates (left) or hd-PS, Basic and Extended Predefined Covariates (right) with Different Aggregation Methods

D. DISCUSSION

We hypothesized that aggregations of medical diagnoses and/or medications into appropriate levels of CCS or ATC would help the performance of the hd-PS, especially with smaller cohort size, rarer outcome incidence or lower exposure prevalence. To explore these hypotheses, we selected a retrospective cohort as the full cohort where, as has been previously observed, the hd-PS adjustment for confounding yielded an adjusted RR substantially close to the RCT findings [18-23] as did propensity score adjustment using a limited number of investigator predefined covariates [13,15,16].

Of the 500 variables identified by hd-PS in the full cohort, most were also identified by hd-PS in the random samples with fewer observations, rarer outcomes, or lower prevalences of treatments. To our knowledge, this is the first study to evaluate the effect of aggregation of medical diagnoses into CCS and/or of medications into ATC on the hd-PS adjustment for confounding in cohorts with small size, rare outcome incidence or low exposure prevalence. Aggregations of medications into ATC level 4 alone or in combination with aggregation of diagnoses into CCS level 1 improved the hd-PS adjustment for confounding in the full cohort and most of the samples. Our results on the effect of aggregating diagnoses is limited, however, by the fact that we did not observe much confounding by co-morbidity in our empirical setting and the little confounding by co-morbidities was in the wrong direction, i.e., away from what we would expect from RCT results.

In general, aggregation of potential covariates into higher-level categories increases the number of covariates that are present in at least 100 observations, the default requirement of the hd-PS, and increases the prevalence of the covariate in exposed and unexposed groups, which increases the covariate's prioritization from the Bross formula if it is associated with treatment [13,26]. But aggregation may simultaneously weaken covariate-exposure and/or covariate-outcome relations, reducing prioritization in the Bross formula. The latter also has the potential to change the impact of control for the aggregated covariate on the adjusted risk ratios.

For example, ICD-9 code 5301 includes 53011 (reflux esophagitis) and the additional codes 53010 (esophagitis unspecified), 53012 (acute esophagitis) and 53019 (other esophagitis). The latter

three codes each occurred in fewer than the 100 observation minimum that hd-PS requires by default and so would not be eligible for inclusion in the propensity score adjustment. With 5-digit granularity for diagnoses, the hd-PS selected ICD-9 code 53011 (frequency 165, covariate-exposure RR=1.3, covariate-outcome RR=5.0 – see Table 5.4). Using 4-digit granularity for diagnoses, the hd-PS selected ICD-9 code 5301 (esophagitis) which had a higher frequency (217) but slightly weaker covariate-exposure (RR=1.2) and covariate-outcome (RR=4.6) associations. Situations like this could account for the slight worsening of confounding control in the 4-digit ICD-9 aggregation compared with the base case (up to 5-digit ICD-9). Additional examples to illustrate the changes in prevalence, covariate-exposure and covariate-outcome relations when we aggregated potential confounders, ICD-9 codes 53011 (reflux esophagitis) and 53081 (esophageal reflux) from 5-digit ICD-9 into 4-, 3-digit ICD-9, and CCS levels 4, 3, 2 and 1 are in Table 5.4. It is worth noting that not all ICD-9 diagnosis codes have their equivalent CCS codes in all 4 levels [68]. This issue was more pronounced in CCS levels 3 and 4. Using the most granular CCS code available for each ICD-9 code in the universal CCS did not improve results in most samples and the full cohort. We also did not observe any benefit while aggregating ICD-9 codes into first 3- or 4-digit categories [67].

Grouping medications into ATC level 4 helped the hd-PS to robustly function in the samples, except for the 20% outcome incidence (category 4). The use of other ATC levels for aggregating medications did not provide benefit and even resulted in some harm. For example, ATC level 4 code B01AC (platelet aggregation inhibitors excluding heparin) includes the following level 5 codes: B01AC04 (clopidrogel), B01AC05 (ticlopidine), B01AC07 (dipyridamole), B01AC23 (cilostazol), and B01AC30 (combined drugs). The latter four codes each occurred in fewer than the 100 observation minimum that hd-PS requires by default and so would not be eligible for inclusion in the propensity score adjustment. With ATC level 5 for medications, the hd-PS selected code B01AC04 (frequency 218, covariate-exposure RR=1.5, covariate-outcome RR=3.8 – Table 5.5). Using ATC level 4 for medications, the hd-PS selected ATC level 4 code B01AC which had a slightly higher frequency (253), the same covariate-exposure (RR=1.5) but slightly weaker covariate-outcome (RR=3.3) associations. Situations like this may account for the observed improvement in confounding control in the ATC level 4 aggregation (e.g., RR of 0.83 in 20% exposure prevalence scenario) compared with

scenarios that used ATC level 5 (e.g., RR of 0.88). Additional examples to illustrate the changes in prevalence, covariate-exposure and covariate-outcome relations from aggregation of clopidrogel and warfarin from level 5 to ATC levels 4, 3, 2 and 1 are in Table 5.5. The ATC level 4 with pharmacological subgroups seems the most appropriate level for aggregation of medications in this study.

It may be argued that our full cohort had relatively low outcome incidence (~0.6%) and only 46 exposed patients with events, so that net benefits even of appropriate aggregation may become smaller in the studies with more common outcomes. Nevertheless, pharmacoepidemiologic studies often have low outcome incidence.

Since CCS has only 18 categories for level 1 and 134 categories for level 2, it could be argued that the benefit from aggregation comes about by enabling more variables from the other data dimensions (medications, inpatient and outpatient procedures) to fit within the 500 variable maximum in the hd-PS default. To address this concern, we also experimented with a maximum of k=3,000 variables and consistently observed the benefit of aggregation of ICD-9 into CCS levels 1 or 2. Similarly, ATC level 1 has 14 groups, whereas level 4 has over 800 categories, but aggregation of medications into ATC level 4 outperformed aggregation into level 1.

Our study has several limitations. Our study used RCT findings [18-23] as expected treatment effect estimates. We thus empirically compared estimates from different aggregations and assumed any treatment effect estimates closer to the RCT findings to be less biased by confounding. Our comparison relies on the assumption that the codes in the original database are accurate. Also, our study is based in a single cohort in which hd-PS performed reasonably well. Fully specified simulations with true risk ratios in diversified scenarios could be used to prove the advantage of aggregation under certain conditions but would be unable to answer the important question of magnitude under realistic assumptions. It is nevertheless unclear whether our findings regarding the effects of aggregation of diagnostic codes and medications on the performance of the hd-PS algorithm apply to other treatment-outcome pairs that may be subject to confounding by different

factors. Studies with few events or small size may have small sample bias or overfit propensity score models and outcome models using propensity score deciles to estimate adjusted risk ratios [79,80] and the small number of UGI complication cases produced imprecise estimates. Finally, the computer time requirements of the hd-PS algorithm constrained our ability to increase the size of our samples beyond 100 for each cohort category. However, each aggregation scenario had six cohort categories (600 samples). Thus, consistent patterns (the combined ATC level 4 plus CCS level 1) are supported by a large number of samples. Users of the hd-PS methodology should screen and remove instrumental variables and collider bias candidates [10-12]. This topic is out of the scope of this study.

Further studies may explore examples of null drug-outcome association, increased drug-outcome risk, more common outcome incidence, to compare the aggregation approaches with the zero-cell correction or exposure-based association selection for the hd-PS [17], to develop appropriate methods to replace missing codes in CCS levels, appropriate aggregations for procedures, simultaneous aggregation of diagnoses, medications and procedures, to evaluate the hd-PS functions in cohorts with different cohort size, outcome incidence and exposure prevalence.

In an empirical pharmacoepidemiologic study using claims data, aggregations of medications into level 4 of the Anatomical Therapeutic Chemical alone or in combination with aggregation of diagnoses into level 1 of the Clinical Classification Software improved the hd-PS adjustment for confounding in most scenarios assessed in an empirical pharmacoepidemiologic example with strong confounding by indication.

Table 5.4. Changes of Prevalences, Covariate-Exposure and Covariate-Outcome Relations When we Aggregated Potential Confounders, ICD-9 Codes 53011 (reflux esophagitis) and 53081 (esophageal reflux) From 5-digit ICD-9 into 4-, 3-digit ICD-9, and Levels 4, 3, 2 and 1 of the Clinical Classification Software (CCS)

Dictionary / Level	Code	Description	Frequency	Frequency Type	Covariate Exposure Risk Ratio	Covariate-Outcome Risk Ratio	Prevalence in both groups	Included in lower level
5-digit ICD-9	53011	REFLUX ESOPHAGITIS	165	once	1.3	5.0	0.01	
5-digit ICD-9	53081	ESOPHAGEAL REFLUX	619	once	1.2	3.4	0.03	
4-digit ICD-9	5301	ESOPHAGITIS	217	once	1.2	4.6	0.01	
	53010	ESOPHAGITIS UNSPECIFIED	<100					No
	53011	REFLUX ESOPHAGITIS	165					Yes
	53012	ACUTE ESOPHAGITIS	<100					No
	53019	OTHER ESOPHAGITIS	<100					No
4-digit ICD-9	5308	OTHER DISORDERS OF ESOPHAGUS	634	once	1.2	3.3	0.03	
	53081	ESOPHAGEAL REFLUX	619					Yes
	53085	BARRETT'S ESOPHAGUS	<100					No
	53089	OTHER DISEASES OF ESOPHAGUS	<100					No
3-digit ICD-9	530	DISEASES OF ESOPHAGUS	827	frequent	1.3	2.4	0.04	
3-digit ICD-9	530	DISEASES OF ESOPHAGUS	827	once	1.2	4.0	0.04	
	5300	ACHALASIA AND CARDIOSPASM	<100					No
	53010	ESOPHAGITIS UNSPECIFIED	<100					Yes
	53011	REFLUX ESOPHAGITIS	165					Yes
	53012	ACUTE ESOPHAGITIS	<100					Yes
	53019	OTHER ESOPHAGITIS	<100					Yes
	5302	ULCER OF ESOPHAGUS	<100					No
	53020	ULCER OF ESOPHAGUS WITHOUT BLEEDING	<100					No
	5303	STRICTURE AND STENOSIS OF ESOPHAGUS	<100					No
	5305	DYSKINESIA OF ESOPHAGUS	<100					No
	5306	DIVERTICULUM OF ESOPHAGUS ACQUIRED	<100					No
	5307	GASTROESOPHAGEAL LACERATION-HEMORRHAGE SYNDROME	<100					No
	53081	ESOPHAGEAL REFLUX	619					Yes
	53085	BARRETT'S ESOPHAGUS	<100					Yes
	53089	OTHER DISEASES OF ESOPHAGUS	<100					Yes
	5309	UNSPECIFIED DISORDER OF ESOPHAGUS	<100					No
CCS Level 4	9.4.1.1	ESOPHAGITIS	217	once	1.2	4.6	0.01	
CCS Level 4	9.4.1.2	OTHER ESOPHAGEAL DISORDERS	673	once	1.2	3.4	0.04	
CCS Level 3	9.4.1	ESOPHAGEAL DISORDERS	828	frequent	1.2	2.4	0.04	
CCS Level 3	9.4.1	ESOPHAGEAL DISORDERS	828	once	1.2	4.0	0.04	
CCS Level 2	9.4	UPPER GASTROINTESTINAL DISORDERS	909	frequent	1.2	2.6	0.05	
CCS Level 2	9.4	UPPER GASTROINTESTINAL DISORDERS	909	once	1.2	4.3	0.05	
CCS Level 1	9	DISEASES OF THE DIGESTIVE SYSTEM	696	frequent	1.1	2.7	0.15	
CCS Level 1	9	DISEASES OF THE DIGESTIVE SYSTEM	2783	once	1.1	2.8	0.15	

Table 5.5. Changes of Prevalences, Covariate-Exposure and Covariate-Outcome Relations When we Aggregated Potential Confounders, Clopidrogel and Warfarin From Level 5 to Levels 4, 3, 2 and 1 of the Anatomical Therapeutic Chemical (ATC) Classification

Dictionary/ Level	Code	Description	Frequency	Frequency Type	Covariate Exposure Risk Ratio	Covariate-Outcome Risk Ratio	Prevalence in both groups	Included in lower level
Generic drug		CLOPIDOGREL	218	once	1.5	3.8	0.012	
Generic drug		CLOPIDOGREL	218	sporadic	1.4	2.9	0.012	
Generic drug		WARFARIN	319	once	1.6	2.0	0.017	
Generic drug		WARFARIN	319	sporadic	1.7	1.3	0.017	
ATC Level 5	B01AC04	CLOPIDOGREL	218	once	1.5	3.8	0.012	
ATC Level 5	B01AC04	CLOPIDOGREL	218	sporadic	1.4	2.9	0.012	
ATC Level 5	B01AA03	WARFARIN	319	once	1.6	2.0	0.017	
ATC Level 5	B01AA03	WARFARIN	319	sporadic	1.7	1.3	0.017	
ATC Level 4	B01AC	PLATELET AGGREGATION INHIBITORS EXCLUDING HEPARIN	253	once	1.5	3.3	0.013	
ATC Level 4	B01AC	PLATELET AGGREGATION INHIBITORS EXCLUDING HEPARIN	253	sporadic	1.5	2.5	0.013	
	B01AC04	CLOPIDOGREL	218					Yes
	B01AC05	TICLOPIDINE	1				0.000	No
	B01AC07	DIPYRIDAMOLE	6				0.000	No
	B01AC23	CILOSTAZOL	25				0.001	No
	B01AC30	COMBINATIONS	11				0.001	No
ATC Level 4	B01AA	VITAMIN K ANTAGONISTS	319	once	1.6	2.0	0.017	
ATC Level 4	B01AA	VITAMIN K ANTAGONISTS	319	sporadic	1.7	1.3	0.017	
	B01AA03	WARFARIN	319					Yes
ATC Level 3	B01A	ANTITHROMBOTIC AGENTS	637	once	1.5	1.5	0.034	
ATC Level 3	B01A	ANTITHROMBOTIC AGENTS	637	sporadic	1.6	2.0	0.034	
ATC Level 2	B01	ANTITHROMBOTIC AGENTS	637	once	1.5	1.5	0.034	
ATC Level 2	B01	ANTITHROMBOTIC AGENTS	637	sporadic	1.6	2.0	0.034	
ATC Level 1	B	BLOOD AND BLOOD FORMING ORGANS	1049	once	1.4	1.4	0.056	
ATC Level 1	B	BLOOD AND BLOOD FORMING ORGANS	1049	sporadic	1.4	1.9	0.056	
ATC Level 1	B	BLOOD AND BLOOD FORMING ORGANS	1049	frequent	1.5	2.0	0.025	

CHAPTER VI

DISCUSSION

A.SUMMARY OF FINDINGS

This dissertation examined some of the factors which can affect the performance of the hd-PS algorithm to adjust confounding for treatment effects using claims databases. The research had three objectives: *1)* To evaluate the performance of the hd-PS algorithm to adjust for confounding of treatment effects in cohorts with different calendar time periods and administrative data sources; *2)* To determine how low outcome incidence or exposure prevalence can degrade hd-PS performance in medium sized or large cohorts; and *3)* To evaluate the effects of aggregating medical diagnoses and/or medications on the hd-PS performance, especially in cohorts with relatively few patients, rare outcome incidence, or low exposure prevalence.

To address the first objective, we used a retrospective cohort of upper gastrointestinal (GI) complications with celecoxib versus nonsteroidal anti-inflammatory drugs (NSAIDs) for osteoarthritis (OA) and rheumatoid arthritis (RA) as an example for assessment of the performance of the hd-PS in the cohorts with multiple settings, since the upper GI complication treatment effect of COX-2 versus NSAIDs is well established based on several Randomized Controlled Trials (RCT) [18-23]. We therefore assumed that a treatment effect estimate closer to 0.5 is less biased by confounding. We used two large claims databases MarketScan and Optum, and created subcohorts before and after 30 September 2004, the date of rofecoxib withdrawal, a drug in the same class of celecoxib, from the US market. Analyses were conducted for individual data sources, subcohorts and the combined two databases. We found that different methods of confounder selection by using expert knowledge only, an automated search via the hd-PS algorithm, and both with the propensity score deciles or greedy

matching inconsistently reduced confounding by indication to estimate the treatment effect for studies with various study periods and administrative data sources.

We did not observe a uniform improvement of confounding control with the hd-PS. In fact we found some settings, especially with propensity score matching, in which the hd-PS performed worse than propensity scores based on predefined covariates. Analyses taking into account various calendar time periods and disparate administrative data sources led to large differences in estimates. The strength of confounding by indication for the effect of non-steroidal anti-inflammatory drugs on upper gastrointestinal complication varied over time before and after the date of voluntary withdrawal of rofecoxib, the same study drug class, from the US market. The hd-PS helped to detect incomplete data of inpatient diagnoses and procedures of certain years of the Optum database as well as programming errors via a very high c-statistics. The hd-PS added little, if anything in our study, compared with the other issues which users of "automated" pharmacoepidemiology in active product safety monitoring systems should be aware of.

To address the second objective, we selected a cohort example of MarketScan, July 2003-September 2004 for resampling. We applied hd-PS to the full study cohort to estimate the treatment effect and used it as the reference value for comparison with results from the generated cohorts. For the 100 samples in each of the cohort categories, we calculated summary statistics for the estimated risk ratios (geometric mean, 25^{th} and 75^{th} percentiles), the mean percentage of covariates selected by hd-PS in the full cohort that were also selected by hd-PS in the samples, the median number of exposed and unexposed subjects, and the median number of exposed and unexposed outcomes. A majority of the covariates that hd-PS identified in the full cohort were also selected by hd-PS in the samples. We found that with >20 exposed outcome events, hd-PS adjusted RRs in the samples were similar to the full cohort values; with ~10 exposed outcome events, the hd-PS functions with less stability; and of the 500 variables identified by hd-PS in the full cohort, most were also identified by hd-PS in the random samples.

To address the third objective, we selected a cohort example of MarketScan, July 2003-September 2004 for resampling studies. In the basic scenario, we applied the hd-PS with up to 5-digit granularity of ICD-9 for inpatient and outpatient diagnoses. We transformed ICD-9 diagnoses into four-level CCS categories via the cross-mapped ICD-9 to CCS multi-level diagnoses table [68]. We separately investigated different levels of ICD-9 granularity by using the first 3- or 4-digit ICD-9 codes. We aggregated medications to five levels of the Anatomical Therapeutic Chemical (ATC) classification of the World Health Organization (WHO) [69]. We found that aggregations of medications into chemical, pharmacological or therapeutic subgroups (level 4) of the Anatomical Therapeutic Chemical classification alone or in combination of aggregation of diagnoses into largest groups (level 1) of the Clinical Classification Software improved the hd-PS adjustment for confounding in studies of treatment effect with small size, rare outcome incidence or infrequent exposure prevalence.

B. PUBLIC HEALTH IMPLICATIONS

To obtain unbiased estimates of drug benefits and harms is important for public health. However, sometimes this information is not available from RCTs. Non-experimental studies try to answer all the questions that remain unanswered by RCTs. Post-marketing drug safety studies often face challenges such as small size, rare incidence of adverse events, or low exposure prevalence after a new drug launch. In addition, active surveillance often generates a large number of safety signals, which emphasizes the need for a method that can rapidly refine a signal. The hd-PS algorithm automatically defines and selects variables for inclusion in the propensity score to adjust treatment effect estimates in studies using healthcare data [13,15,16]. The hd-PS algorithm could reduce programming time and error, and run in studies pooling multiple claims databases [13,14]. The hd-PS performance has been evaluated with few outcome events or few exposed subjects in small cohorts only [17]. Prior studies demonstrated the hd-PS was the potential algorithm software for active drug safety monitoring systems using longitudinal healthcare claims databases [14]. It is a promising algorithm for studies using healthcare claims data. However, it is not known whether different calendar time periods, data sources, low outcome incidence or exposure prevalence can

degrade hd-PS performance in medium sized or large cohorts. Also, no study to date has assessed how hd-PS performance is affected by aggregating medical diagnoses and/or medications, especially in cohorts with relatively few patients, rare outcome incidence, or low exposure prevalence. Extensive testing of the performance of hd-PS in multiple settings will provide information on factors affecting the hd-PS performance and to determine the value of this new approach. Finally, solutions to improve the hd-PS performance particularly in specific settings will contribute to the research community for active drug safety surveillance.

The results of this dissertation research have several implications for public health and clinical practice. First, the finding provides a more accurate picture of potential factors that can affect the performance of the hd-PS algorithm to adjust for confounding of treatment effects in healthcare claims data. Users of the hd-PS should be aware of these potential factors such as calendar time periods, administrative data sources, cohort size, low outcome incidence and exposure prevalence. Three different methods of confounder selection by using expert knowledge only, an automated search via the hd-PS algorithm, and both, inconsistently reduced confounding by indication. No method proved better than others in cohorts with different calendar time periods, administrative data sources or using propensity score deciles or 1:1 greedy matching. During the study implementation, another potential benefit of the hd-PS is to help to detect incomplete data of inpatient diagnoses and procedures of certain years of the Optum database as well as programming errors via very high c-statistics. Second, the finding of our study provides evidence that not only small size, but low outcome incidence or exposure prevalence also degrade hd-PS performance in medium sized or large cohorts.

Lastly, this is the first study to evaluate the effect of aggregation of medical diagnoses into CCS and/or of medications into ATC on the hd-PS adjustment for confounding in cohorts with small size, rare outcome incidence or low exposure prevalence. Aggregations of medications into chemical, pharmacological or therapeutic subgroups (level 4) of the Anatomical Therapeutic Chemical classification alone or in combination of aggregation of diagnoses into systemic groups (level 1) of the Clinical Classification Software improved the hd-PS adjustment for confounding in studies of

treatment effect with small size, rare outcome incidence or infrequent exposure prevalence. These findings will provide alternative solutions to improve hd-PS performance in cohorts with small size, rare outcome incidence or low exposure prevalence. It is important to select appropriate levels of ATC and CCS for aggregations. In general, aggregation of potential covariates into higher-level categories increases the number of covariates that are present in at least 100 observations, the default requirement of the hd-PS, and increases the prevalence of the covariate in exposed and unexposed groups, which increases the covariate's prioritization from the Bross formula if it is associated with treatment. But aggregation may simultaneously weaken covariate-exposure and/or covariate-outcome relations, reducing prioritization in the Bross formula. The latter also has the potential to change the impact of control for the aggregated covariate on the adjusted risk ratios.

C. STRENGTHS

First, as per our best knowledge, this is the first study to evaluate the effect of aggregation of medical diagnoses into CCS and/or of medications into ATC on the hd-PS adjustment for confounding in cohorts with small sizes, rare outcome incidence or infrequent exposure prevalence. In an empirical pharmacoepidemiologic study using claims data, aggregations of medications into level 4 of the Anatomical Therapeutic Chemical alone or in combination with aggregation of diagnoses into level 1 of the Clinical Classification Software improved the hd-PS adjustment for confounding in most scenarios assessed in studies of treatment effect with small size, rare outcome incidence or infrequent exposure prevalence. This can be explained by the fact that aggregations of medical diagnoses or medications into their certain levels of ATC or CCS increased their respective prevalence, minimized the missing covariate-exposure or covariate-outcome associations for the hd-PS adjustment.

Second, our research was the first to address the low outcome incidence or exposure prevalence can degrade hd-PS performance in medium sized or large cohorts. When sampling smaller studies, rarer outcomes, or lower prevalences of treatments, the mean average percentage of the hd-PS covariates used in the full cohort, also identified in the samples decreased. Its frequency

was not only dependent on cohort size, numbers of exposed subjects or outcomes [17], but also outcome incidence or exposure prevalence. In other words, even with the same cohort size and nearly the same number of exposed outcomes, frequency of hd-PS variables identified in the samples was lower in samples with rare outcome incidence than in samples with less frequent exposure. This is explicitly demonstrated by the fact that the hd-PS at default setting requires a variable with at least 100 frequencies in combined exposed and unexposed groups, and nonmissing covariate-exposure or covariate-outcome associations to be eligible to retain in the variable selection for hd-PS adjustment [13, 26].

Third, our sampling techniques which did not affect the treatment effect estimate while being able to keep total sample size constant for outcome incidence or exposure prevalence samplings, were validated in 1,000 simulations (Appendix B).

Finally, we used two large longitudinal healthcare claims databases to explore the impact of calendar time periods and disparate administrative data sources on the performance of the hd-PS to estimate treatment effect. There were a large number of patients available to study with the potential to look at longer periods of pre-diagnosis times that would be found in other databases. This large sample also allowed sufficient power to examine low frequency of upper GI complications.

D. LIMITATIONS

Our study has several limitations. First, we used randomized clinical trial findings [18-23] as the expected treatment effect estimates. We thus empirically compared estimates from different aggregations and assumed any treatment effect estimates closer to the randomized clinical trial findings to be less biased by confounding. Fully specified simulations with true risk ratios in diversified scenarios could be explored instead of resampling of the real-world cohort. Second, it is unclear whether our findings regarding the hd-PS algorithm apply to other treatment-outcome pairs that may be subject to confounding by different factors. Third, studies with few events or small size may have small sample bias or overfit propensity score and outcome models [79, 80]. Fourth, the small number

of upper gastrointestinal complication cases produced a wide 95% confidence interval. Fifth, the Optum database did not record inpatient diagnoses and inpatient procedures which occurred before the year 2003, thus it was limited to creating a bigger and more comparable cohort with the study drug initiation before 30 September 2004. Finally, there were the potential confounding by severity of RA or OA, confounding by indication and healthcare factors; the lack of validation of exposure and outcome measurement; and the potential for information bias resulting from varying records and approaches of specialists, emergency room or hospitalization visits.

E. FUTURE RESEARCH

Further studies may explore examples of null drug-outcome association, increased drug-outcome risk, more common outcome incidence, to compare the aggregation approaches with the zero-cell correction or exposure-based association selection for the hd-PS [17], to develop appropriate methods to replace missing codes in CCS levels, appropriate aggregations for procedures, simultaneous aggregation of diagnoses, medications and procedures into the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) to evaluate the hd-PS functions in cohorts with different cohort size, outcome incidence and exposure prevalence. In this example, we only used the outcome model without trimming of the PS distribution to maintain the cohort size relatively constant at each sampling rate for size sampling and for outcome incidence and exposure prevalence samplings. Notwithstanding, future study on the performance of the hd-PS with propensity score matching, one-sided or two-sided trimming or different trimming levels of propensity score distribution may be carried out. Using richer information databases such as electronic medical record data to minimize unmeasured confounders should be also considered in future studies in order to further investigate the role of aggregation methods for the hd-PS algorithm and automated confounding control [17]. Finally, future studies should validate internal assumptions of sufficient for controlling confounding with identified measured covariates of the hd-PS algorithm [82] or alternative variable selection strategy based on disease risk scores for multiple settings [83].

F. CONCLUSIONS

Different methods of confounder selection by using expert knowledge only, an automated search via the hd-PS algorithm, and both with the propensity score deciles or greedy matching inconsistently reduced confounding by indication to estimate the treatment effect for studies with various calendar time periods and administrative data sources. Users should be aware of potential factors affecting the hd-PS performance such as calendar time periods, data sources, small cohort low size, outcome incidence and low exposure prevalence. Users should also be aware that the confounding control achieved by the hdPS is not always at least as good as the one obtained by selecting covariates based on expert knowledge. In an example of pharmacoepidemiologic studies using claims data, aggregations of medications into chemical, pharmacological or therapeutic subgroups (level 4) of the Anatomical Therapeutic Chemical classification alone or in combination of aggregation of diagnoses into largest groups (level 1) of the Clinical Classification Software improved the hd-PS adjustment for confounding in most scenarios including ones with small cohort size, rare outcome incidence, and low exposure prevalence.

Table A.1. Examples of mappings from ICD-9 diagnoses into the multi-level Clinical Classification Software (CCS)

ICD-9	ICD-9 _TERM	LEVEL 1	LEVEL 1 LABEL	LEVEL 2	LEVEL 2 LABEL	LEVEL 3	LEVEL 3 LABEL	LEVEL 4	LEVEL 4 LABEL
535	Gastritis and duodenitis	9	Diseases of the digestive system	9.4	Upper gastrointestinal disorders	9.4.3	Gastritis and duodenitis		
5350	Acute gastritis	9	Diseases of the digestive system	9.4	Upper gastrointestinal disorders	9.4.3	Gastritis and duodenitis	9.4.3.1	Acute gastritis
53500	Acute gastritis, without mention of hemorrhage	9	Diseases of the digestive system	9.4	Upper gastrointestinal disorders	9.4.3	Gastritis and duodenitis	9.4.3.1	Acute gastritis
53501	Acute gastritis, with hemorrhage	9	Diseases of the digestive system	9.4	Upper gastrointestinal disorders	9.4.3	Gastritis and duodenitis	9.4.3.1	Acute gastritis
5351	Atrophic gastritis	9	Diseases of the digestive system	9.4	Upper gastrointestinal disorders	9.4.3	Gastritis and duodenitis	9.4.3.2	Other specified gastritis
53510	Atrophic gastritis, without mention of hemorrhage	9	Diseases of the digestive system	9.4	Upper gastrointestinal disorders	9.4.3	Gastritis and duodenitis	9.4.3.2	Other specified gastritis
53511	Atrophic gastritis, with hemorrhage	9	Diseases of the digestive system	9.4	Upper gastrointestinal disorders	9.4.3	Gastritis and duodenitis	9.4.3.2	Other specified gastritis

APPENDIX B. SIMULATION RESULTS OF OUTCOME AND EXPOSURE SAMPLINGS

Table B.1. Simulation Results of Outcome Sampling With Recoded Cases: Constant Adjusted Treatment Effect Estimates (True Treatment Effect =0.5) and Cohort Sizes

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
N	1000	10000.00	0	10000.00	10000.00	10000.00
pe	1000	33.1584500	0.4738969	33.1500000	31.3400000	34.7500000
iy	1000	9.5184200	0.2950972	9.5300000	8.6800000	10.4200000
RRect	1000	0.8961194	0.0633242	0.8959158	0.7027671	1.1072531
RReat	1000	**0.5019544**	0.0367811	0.5013105	0.3893997	0.6357730
N50	1000	**10000.00**	0	10000.00	10000.00	10000.00
N20	1000	**10000.00**	0	10000.00	10000.00	10000.00
N10	1000	**10000.00**	0	10000.00	10000.00	10000.00
N5	1000	**10000.00**	0	10000.00	10000.00	10000.00
pey50	1000	33.1584500	0.4738969	33.1500000	31.3400000	34.7500000
iyy50	1000	4.7617000	0.1476405	4.7700000	4.3400000	5.2100000
pey20	1000	33.1584500	0.4738969	33.1500000	31.3400000	34.7500000
iyy20	1000	1.9078600	0.0590110	1.9100000	1.7400000	2.0900000
pey10	1000	33.1584500	0.4738969	33.1500000	31.3400000	34.7500000
iyy10	1000	0.9564800	0.0295886	0.9600000	0.8700000	1.0500000
pey5	1000	33.1584500	0.4738969	33.1500000	31.3400000	34.7500000
iyy5	1000	0.4806500	0.0149633	0.4800000	0.4400000	0.5300000
RRecy50	1000	0.9014638	0.0921339	0.8980028	0.6697348	1.1971396
RReay50	1000	**0.5054475**	0.0545001	**0.5041009**	0.3488935	0.6935752
RRecy20	1000	0.9043103	0.1507262	0.8985488	0.5032395	1.5996958
RReay20	1000	**0.5081023**	0.0889443	**0.5050795**	0.2569943	0.9364096
RRecy10	1000	0.9053591	0.2126796	0.8786616	0.4125483	1.6620532
RReay10	1000	**0.5083839**	0.1257661	**0.4952604**	0.2229691	1.0053211
RRecy5	1000	0.9129616	0.3122159	0.8722794	0.3164708	2.4314710
RReay5	1000	**0.5159485**	0.1904948	**0.4885627**	0.1530384	1.4216319

Note: 50, 20, 10 and 5 mean 50%, 20%, 10% and 5% sampling rate
N = number of subjects in the full cohort
pe = exposure prevalence in the full cohort
iy = outcome incidence in the full cohort
RRect = crude risk ratio (RR) in the full cohort
RReat = adjusted RR in the full cohort

N50 = number of subjects in 50% samples
pey50 = exposure prevalence in 50% samples
iyy50 = outcome incidence in 50% samples
RRecy50= crude RR in 50% samples
RReay50= adjusted RR in 50% samples

Table B.2. Simulation Results of Exposure Sampling With Replacement of Unexposed: Constant Adjusted Treatment Effect Estimates (True Treatment Effect =0.5) and Cohort Sizes

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
N	1000	**10000.00**	0	10000.00	10000.00	10000.00
pe	1000	33.1584500	0.4738969	33.1500000	31.3400000	34.7500000
iy	1000	9.5184200	0.2950972	9.5300000	8.6800000	10.4200000
RRect	1000	0.8961194	0.0633242	0.8959158	0.7027671	1.1072531
RReat	1000	0.5019544	0.0367811	0.5013105	0.3893997	0.6357730
NRe50	1000	**10000.00**	0	10000.00	10000.00	10000.00
NRe20	1000	**10000.00**	0	10000.00	10000.00	10000.00
NRe10	1000	**10000.00**	0	10000.00	10000.00	10000.00
NRe5	1000	**10000.00**	0	10000.00	10000.00	10000.00
pe50	1000	16.5816100	0.2368814	16.5800000	15.6700000	17.3800000
iy50	1000	9.6494500	0.3373354	9.6500000	8.6500000	10.6300000
pe20	1000	6.6356600	0.0947520	6.6300000	6.2700000	6.9500000
iy20	1000	9.7432500	0.3600978	9.7500000	8.5300000	10.8500000
pe10	1000	3.3202800	0.0473853	3.3200000	3.1400000	3.4800000
iy10	1000	9.7587400	0.3621499	9.7600000	8.5600000	10.8700000
pe5	1000	1.6627000	0.0237671	1.6600000	1.5700000	1.7400000
iy5	1000	9.7794000	0.3765286	9.7800000	8.6500000	11.0800000
RRec50	1000	0.8951330	0.0808613	0.8947324	0.6571889	1.1515775
RRea50	1000	**0.5014691**	0.0465928	0.5001284	0.3600029	0.6620642
RRec20	1000	0.8962348	0.1269376	0.8888930	0.5581104	1.2629862
RRea20	1000	**0.5022340**	0.0701844	0.5001651	0.3035006	0.7344370
RRec10	1000	0.9003910	0.1653225	0.8989558	0.4405790	1.4197919
RRea10	1000	**0.5039989**	0.0896516	0.5029586	0.2638169	0.8039307
RRec5	1000	0.8894906	0.2336480	0.8790189	0.2562333	1.8299925
RRea5	1000	**0.4983340**	0.1283653	0.4903760	0.1507900	0.9313329

Note: 50, 20, 10 and 5 mean 50%, 20%, 10% and 5% sampling rate
N = number of subjects in the full cohort
pe = exposure prevalence in the full cohort
iy = outcome incidence in the full cohort
RRect = crude risk ratio (RR) in the full cohort
RReat = adjusted RR in the full cohort

N50 = number of subjects in 50% samples
pe50 = exposure prevalence in 50% samples
iy50 = outcome incidence in 50% samples
RRect50= crude RR in 50% samples
RReat50= adjusted RR in 50% samples

REFERENCES

1. Davis RL, Kolczak M, Lewis E, et al. Active surveillance of vaccine safety: a system to detect early signs of adverse events. *Epidemiology* 2005; 16(3): 336-341.

2. 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*. 2007; 16(12): 1275-1284.

3. Lieu TA, Kulldorff M, Davis RL, et al. Real-Time vaccine safety surveillance for the early detection of adverse events. *Med Care* 2007; 45(10): S89-S95.

4. Rubin DB. Estimating causal effects from large data sets using the propensity score. *Ann Intern Med* 1997; 127:757–63.

5. Stürmer T, Schneeweiss S, Brookhart MA, et al. 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.* 2005*;* 161 *(*9*):*891-898.

6. Stürmer T, Joshi M, Glynn RJ, et al. A review of applications of propensity score methods showed increased use but infrequently different estimates compared with other methods. *J Clin Epidemiol.* 2006; 59: 437–447.

7. Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology*. Basic Clin Pharmacol Toxicol.* 2006*;* 98*:*253-9*.*

8. Perkins SM, Tu W, Underhill MG, et al. The use of propensity scores in pharmacoepidemiologic research. *Phamacoepidemiol Drug Saf*. 2000; 9:93–101.

9. Robins JM, Mark SD, Newey WK. Estimating exposure effects by modeling the expectation of exposure conditional on confounders. *Biometrics* 1992; 48:479–95.

10. Rubin DB. On principles for modeling propensity score in medical research. *Phamacoepidemiol Drug Saf.* 2005; 14:227–238.

11. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol*. 2006 Jun 15; 163(12):1149-56.

12. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology* 2003;14(3):300-306*.*

13. Schneeweiss S, Rassen JA, Glynn RJ, et al. High-Dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data. *Epidemiology.* 2009 Jul; 20(4):512-22.

14. Rassen AJ, Schneeweiss S. Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system. *Pharmacoepidemiology and Drug Safety*. Supplement: The U.S. Food and Drug Administration's Mini-Sentinel Program. 2012; 21(S1), 41–49.

15. Toh S, Rodríguez AGL, Hernán AM. Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records. *Pharmacoepidemiology and drug safety* 2011; 20: 849–857.

16. Schneeweiss S, Rassen J. Pharmacoepidemiol Drug Saf. 2011 Oct;20(10):1110-1; author reply 1112. doi: 10.1002/pds.2238. PMID: 21953847

17. Rassen AJ, Glynn JR, Brookhart MA, et al. Covariate Selection in High-Dimensional Propensity Score Analyses of Treatment Effects in Small Samples. *American Journal of Epidemiology* 2011; 173 (12), 1404-1413.

18. Silverstein FE, Faich G, Goldstein JL, et al. Gastrointestinal toxicity with celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis: the CLASS study: A randomized controlled trial. Celecoxib Long-term Arthritis Safety Study. *JAMA*. 2000 Sep 13; 284(10):1247-55.

19. Singh G, Fort JG, Goldstein JL, et al. Celecoxib versus naproxen and diclofenac in osteoarthritis patients: SUCCESS-I Study. *Am J Med.* 2006 Mar;119(3):255-66.

20. Deeks JJ, Smith LA, Bradley MD. Efficacy, tolerability, and upper gastrointestinal safety of celecoxib for treatment of osteoarthritis and rheumatoid arthritis: systematic review of randomised controlled trials. *BMJ*. 2002 Sep 21;325 (7365):619.

21. Moore RA, Derry S, Makinson GT, et al. Tolerability and adverse events in clinical trials of celecoxib in osteoarthritis and rheumatoid arthritis: systematic review and meta-analysis of information from company clinical trial reports. *Arthritis Res Ther.* 2005;7(3):R644-65.

22. Goldstein JL, Silverstein FE, Agrawal NM, et al. Reduced risk of upper gastrointestinal ulcer complications with celecoxib, a novel COX-2 inhibitor, *American Journal of Gastroenterology* 2000; 95 (7):1681-1690.

23. Goldstein JL. Significant upper gastrointestinal events associated with conventional NSAID versus celecoxib. *J Rheumatol.* Suppl. 2000 Oct; 60:25-8.

24. FDA Public Health Advisory: Safety of Vioxx. (accessed 21 Nov 2011 at http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm106274.htm)

25. Mack D. Christina, Robert Glynn. Stürmer T, Time-varying propensity scores and hazard ratio estimation. *Pharmacoepidemiology and Drug Safety* 2011; Volume 20, S 128, Abstract 296.

26. Bross ID. Spurious effects from an extraneous variable. *J Chronic Dis.* 1966; 19(6):637–647.

27. Piazza-Hepp TD, Kennedy DL. Reporting of adverse events *to MedWatch. Am J Health Syst Pharm* 1995; 52(13): 1436-1439.

28. Brewer T, Colditz GA. Postmarketing surveil*lance and adverse dru*g reactions: current perspectives and future needs. *JAMA* 1999; 281(9): 824-829.

29. Wysowski DK, Swartz L. Adverse drug event *surveillance and drug* withdrawals in the United States, 1969-2002: the imp*ortan*ce of reporting suspected reactions. *Arch Intern Med* 2005 165(12): 1363-1369.

30. Furberg CD, Levin AA, Gross PA, Shapiro RS, Strom BL. The FDA and Drug Safety. A Proposal for Sweeping *Changes Arch Intern Med.* 2006**;** 166:1938-1942.

31. Psaty BM, Korn D. Congress responds to the IOM drug safety report-in full. *JAMA* 2007; 298(18): 2185-2187.

32. Chan KA, Hauben M. Signal detection in pharmacovigilance: empirical evaluation of data mining tools**.** *Pharmacoepidemiology and Drug Safety.* Volume 14, Issue 9, Date: September 2005, Pages: 597-599

33. Platt R, Madre L, Reynolds R, Tilson H. Active drug safety surveillance: a tool to improve public health. *Pharmacoepidemiology and Drug Safety*. Volume 17, Issue 12, Date: December 2008, Pages: 1175-1182.

34. Mann RD, Wilton LV, Pearce GL, Mackay FJ, Dunn NR. Prescription-event monitoring (PEM) in 1996 - a method of non-interventional observational cohort pharmacovigilance. *Pharmacoepidemiology and Drug Safety.* Volume 6, Issue S3, Date: October 1997, Pages: S5-S11. R. D.

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

36. Brown JS, Kulldorff M, Petronis KR, Reynolds R, Chan KA, Davis RL, Graham D, Andrade SE, Raebel MA, Herrinton L, Roblin D, Boudreau D, Smith D, Gurwitz JH, Gunter MJ, Platt R. Early adverse drug event signal detection within population-based health networks using sequential methods: key methodologic considerations. *Pharmacoepidemiol Drug Saf.* 2009 Mar; 18(3):226-34.

37. Alina Baciu KS, Sheila P B. The Future of Drug Safety: Promoting and Protecting the Health of the Public*. Institute of Medicine of the National Academies*: Washington, D.C., 2006; Committee on the Assessment of the US Drug Safety System.

38. McClellan M. Drug safety reform at the FDA-pendulum swing or systematic improvement? *N Engl J Med.* 2007; 356(17): 1700-1702.

39. Psaty BM, Charo RA. FDA responds to institute of medicine drug safety recommendations-in part. *JAMA* 2007; 297(17): 1917-1920

40. Bright RA, Nelson RC. Automated support for pharmacovigilance: a proposed system Pharmacoepidemiology and Drug Safety. Volume 11, Issue 2, Date: March 2002, Pages: 121-125

41. US Food and Drug Administration. The Sentinel Initiative: National Strategy for Monitoring Medical Product Safety. (accessed 20 June 2011 at http://www.fda.gov/oc/initiatives/advance/reports/report0508.html)

42. Anderson RM. Revisiting the behavioral model and access to medical care: Does it matter? *J Health Soc Behav.* 1995;36:1–10.

43. Schneeweiss S, Glynn RJ, Avorn J, Solomon DH. A Medicare database review found that physician preferences increasingly outweighed patient characteristics as determinants of first-time prescriptions for COX-2 inhibitors. *J Clin Epidemiol.* 2005; 58:98–102.

44. Roblin DW, Platt R, Goodman MJ, et al. Effect of increased cost-sharing on oral hypoglycemic use in five managed care organizations: how much is too much? *Med Care.* 2005;43:951–959

45. Seeger JD, Kurth T, Walker AM. Use of propensity score technique to account for exposure-related covariates: an example and lesson. *Med Care*. 2007;45:S143–S148.

46. Johannes CB, Koro CE, Quinn SG, et al. The risk of coronary heart disease in type 2 diabetic patients exposed to thiazolidinediones compared to metformin and sulfonylurea therapy. *Pharmacoepidemiol Drug Saf*. 2007;16:504–512.

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

48. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc*. 1984;79(387):516–524.

49. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat.* 1985;39(1):33–38.

50. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11(5):550–560

51. 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.* 2006;163(3):262–270.

52. Seeger JD, Williams PL, Walker AM. An application of propensity score matching using claims data. *Pharmacoepidemiol Drug Saf.* 2005;14:465–476.

53. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Serv Outcomes Res Methodol*. 2001;2:259 –278.

54. Rubin DB. Estimating causal effects from large data sets using the propensity score. *Ann Intern Med* 1997;127:757–63.

55. Brookhart MA, Stürmer T, Glynn RJ, et al. Confounding Control in Healthcare Database Research: Challenges and Potential Approaches. *Med Care.* 2010; 48 (6): S114-S120.

56. Greenland S. Invited commentary: Variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol*. 2008; 167:523–529.

57. Le VH**,** Beach JK, Powell G, et al. Performance of a semi-automated approach for risk estimation using a common data model for longitudinal healthcare databases. *Statistical Methods in Medical Research* 2011, 0(0) 1–17.

58. Schneeweiss, S., Gagne JJ, Glynn RJ, Ruhl M, Rassen JA. Assessing the comparative effectiveness of newly marketed medications: methodological challenges and implications for drug development. *Clin Pharmacol Thera.* 2011 Dec;90(6):777-90. doi: 10.1038/clpt.2011.235. Epub 2011 Nov 2.

59. Schneeweiss S, Avorn, J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J. Clin. Epidemiol.* 2005, 58, 323–337

60. Walker AM. Confounding by indication. *Epidemiology* 1996;7: 335–6.

61. Schneeweiss, S., Glynn, R.J., Avorn, J. & Solomon, D.H. A Medicare database review found that physician preferences increasingly outweighed patient characteristics as determinants of first-time prescriptions for COX-2 inhibitors. *J. Clin. Epidemiol.2005;* 58, 98–102.

62. Hennessy S, Bilker WB,Weber A, Strom BL. Descriptive analyses of the integrity of a US Medicaid claims database. *Pharmacoepidemiol Drug Saf* 2003;12:103–11.

63. Rassen AJ, Avorn J, Schneeweiss S. Multivariate-adjusted pharmacoepidemiologic analyses of confidential information pooled from multiple health care utilization databases. *Pharmacoepidemiol Drug Saf*. 2010 August; 19(8): 848–857.

64. Parsons LS. *Reducing bias in a propensity score matched-pair sample using greedy matching techniques.* 2001(accessed 1 November 2011 at www2.sas.com/proceedings/sugi26/p214-26.pdf)

65. MarketScan® Commercial Claims and Encounters of Thomson Reuters Healthcare. (accessed 21 November 2011 at http://thomsonreuters.com/products_services/healthcare/healthcare_products/a-z/marketscan_research_analytics/).

66. Optum® Impact® National Managed Care Benchmark Database (OptumInsight, Eden Prairie, MN) (accessed 21 November 2011 at http://www.i3global.com/Home/).

67. Hoskins H Jr, Hildebrand P, Lum F. The American Academy of Ophthalmology Adopts SNOMED CT as Its Official Clinical Terminology. *Ophthalmology*, Volume 115, Issue 2, 225-226.

68. Clinical Classifications Software (CCS) for ICD-9-CM. (accessed April 20, 2012 at http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp).

69. The Anatomical Therapeutic Chemical (ATC) classification. (accessed April 20, 2012 at http://www.whocc.no/).

70. Fogerty MD, Naji NN, Nanney L, Arbogast PG, Poulose BB, Barbul A. Risk factors for pressure ulcers in acute care hospitals. Wound Repair & Regeneration. 16(1):11-18, January/February 2008.

71. Radley D C; Gottlieb DJ; Fisher ES; Tosteson Anna N A. Comorbidity risk-adjustment strategies are comparable among persons with hip fracture. *Journal of clinical epidemiology* 2008;61(6):580-7.

72. Raiford DS, Perez Gutthann S, Garcia Rodriguez LA. Positive predictive value of ICD-9 codes in the identification of cases of complicated peptic ulcer disease in the Saskatchewan hospital automated database. *Epidemiology.* 1996;7: 101–104.

73. García Rodríguez LA, Jick H. Risk of upper gastrointestinal bleeding and perforation associated with individual non-steroidal anti-inflammatory drugs. *Lancet* 1994; 343: 769–772.

74. Gutthann SP, García Rodríguez LA, Raiford DS. Individual nonsteroidal antiinflammatory drugs and other risk factors for upper gastrointestinal bleeding and perforation. *Epidemiology* 1997; 8: 18–24.

75. Hernández-Díaz S, García Rodríguez LA. Association between nonsteroidal anti-inflammatory drugs and upper gastrointestinal tract bleeding/perforation: an overview of epidemiologic studies published in the 1990s. *Arch Intern Med* 2000; 160: 2093–2099.

76. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol*. 1987; 125(5):761-768.

77. Greenland S, Neutra R. Control of confounding in the assessment of medical technology. *Int J Epidemiol.* 1980 Dec; 9(4):361-7.

78. Brookhart MA, Wang PS, Solomon DH, et al. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology.* 2006 May; 17(3):268-75.

79. Greenland S, Schwartzbaum JA, Finkle WD. Problems due to small samples and sparse data in conditional logistic regression analysis. *Am J Epidemiol*. 2000;151 (5):531–539.

80. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996; 49(12):1373–1379.

81. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol. 2*003;158 (9): 915–920.

82. Joffe MM. Exhaustion, automation, theory, and confounding. *Epidemiology.* 2009 Jul;20*(4):512-22.*

83. *Gly*nn RJ, Gagne JJ, Schneeweiss S. Role of disease risk scores in comparative effec*tiveness rese*arch with emerging therapies. *Pharmacoepidemiol Drug Saf.* 2012 May;21 Suppl 2:138-47. doi: 10.1002/pds.3231.