MINI-SENTINEL METHODS

STATISTICAL METHODS FOR IMPROVING CONFOUNDER ADJUSTMENT FOR EMERGENT TREATMENT COMPARISON

Prepared by: Stanley Xu, PhD,[1] Susan Shetterly, MS,[1] Andrea J. Cook, PhD,[2] Marsha A. Raebel, PharmD,[1] Sunali Goonesekera, MS,[3] Azadeh Shoaibi, MS, MHS,[4] Eric Frimpong, PhD,[4] Jason Roy, PhD,[5] Bruce Fireman, MS,[6] Brad McEvoy, PhD[4]

Author Affiliations: 1. The Institute for Health Research, Kaiser Permanente Colorado, Denver, CO; 2. Biostatistics Unit, Group Health Research Institute, Seattle, WA; 3. Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA; 4. Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD; 5. Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA; 6. Division of Research, Kaiser Permanente Northern California, Oakland, CA.

February 16, 2014

Mini-Sentinel is a pilot project sponsored by the U.S. Food and Drug Administration (FDA) to inform and facilitate development of a fully operational active surveillance system, the Sentinel System, for monitoring the safety of FDA-regulated medical products. Mini-Sentinel is one piece of the Sentinel Initiative, a multi-faceted effort by the FDA to develop a national electronic system that will complement existing methods of safety surveillance. Mini-Sentinel Collaborators include Data and Academic Partners that provide access to health care data and ongoing scientific, technical, methodological, and organizational expertise. The Mini-Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223200910006I.

Mini-Sentinel Methods

Statistical Methods for Improving Confounder Adjustment For Emergent Treatment Comparison

Table of Contents

I. INTRODUCTION

A. BACKGROUND

Appropriate confounder adjustment is critical in postmarket surveillance because patients are not randomized as in clinical trials. Monitoring newly marketed treatments is particularly challenging due to 1) low uptake of new treatments in the early years of treatment availability; 2) rare occurrence of adverse events (related to low treatment uptake or to the rare nature of the adverse event); and 3) patients treated early during follow-up being different from those treated later (e.g., critically ill patients who do not respond to the standard treatment may be selectively prescribed the new treatment when it is first marketed). Propensity Score (PS) techniques may not be optimal when low use predominates and may require modification as usage increases. Even though Disease Risk Score (DRS) techniques may be advantageous in such scenarios, these too may pose problems when adverse outcomes are rare and/or when potential confounders are strongly associated with treatment.

For this project, we evaluate different statistical choices for confounder adjustment, namely PS and DRS methods and regression, to develop heuristics for confounder adjustment for emerging therapies, and consider expert opinion, feasibility, and results from simulations and literature review to guide our decisions. We then use Mini-Sentinel (M-S) data resources to evaluate implementation of these confounder adjustment methods in a sequential testing framework and provide recommendations for improving confounding adjustment for emergent treatment comparison.

B. SPECIFIC AIMS

This project aims to explore how and whether PS and DRS methods and regression could be implemented to adequately adjust for multiple confounders when evaluating adverse events associated with newly marketed treatments in a sequential testing framework.

II. SIMULATIONS

A. INTRODUCTION/OVERVIEW

We performed simulations to identify optimal confounder adjustment methods under various scenarios in which the frequency of treatment and outcome ranged from rare to common. We then used these results to inform selection of appropriate analytic methods for a methodological example. Incident users in each year of follow-up were used in this simulation study. Disease characteristics and demographic characteristics were measured at entry and were included in propensity score models and disease risk score models. Individuals were not followed longitudinally. For cumulative data runs, individuals from earlier years retain their original covariate values. Their data is used to support models of subsequent years but only new, incident users of each study year are given scores based on these new models.

B. METHODS

1. Simulation of Treatment

We simulated whether an individual receives the treatment of interest (Z=1) or comparator treatment (Z=0) using a logistic regression model (1). We included age (centered at 50 years) in the model as a

continuous variable (x_1), gender (x_2), presence of acute disease (x_3), presence of chronic disease (x_4), and six other covariates (x_5,…, x_{10}) as binary variables. Exact distributions of variables are specified in Section B.8. We defined follow-up time (t for year) as discrete times 1 to 10. For each year, there was a new cohort of patients. The follow-up time (t) was included in the model (1) as a continuous variable.

$$\text{Logit}\,(Z = 1|x_1, x_2, x_3 \ldots) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{10} x_{10} + \beta_{11} t + \beta_{12} x_4 t \quad (1)$$

In the above model, β_0 is the intercept, β_1, β_2 , …, β_{10} are coefficients for confounders $x_1, \ldots, x_{10}$, and β_{11} is the coefficient for time t. Inclusion of the term $\beta_{11} t$ allows the uptake of a new treatment to increase over time steadily. We will also simulate the situation where sicker or healthier individuals may tend to receive the treatment of interest earlier by including the interactions between chronic disease (x_4) and time (t).

2. Simulation of Outcome

At each of the 10 discrete times (t), we simulated the outcome (Y=0 or 1) using model (2).

$$\text{Logit}\,(Y = 1|\text{Z}, x_1, x_2, x_3 \ldots) = \theta_0 + \theta_Z Z + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots + \theta_{10} x_{10} \quad (2)$$

In the above outcome model,θ_0 is the intercept, and θ_1,… , θ_{10} are the coefficients for confounders x_1,…,x_{10}. The parameter for the treatment (θ_Z) represents the log odds ratio of having the outcome for those taking the drug of interest compared to those on the active comparator drug after adjusting for all variables in the model. When the probability of occurrence of adverse events is less than 5%, the odds ratio exp(θ_Z) is approximately the relative risk.

3. PS and DRS Models

The idea of active surveillance is to repeatedly over time assess for the existence of elevated risk of a given outcome due to treatment of interest. For example we may plan to look at the data quarterly and if we do not find evidence for an elevated risk at a given quarter we continue monitoring (e.g., for up to 3 years), but if we do find elevated risk we may signal or alert to conduct a more thorough investigation of the data. To do this in a sequential framework we must incorporate differential confounding over time and handle the sequential testing in which at earlier looks we have estimated a risk that did not indicate a signal and decided to continue monitoring. To properly handle such data we form a sequential monitoring framework that will be discussed in more detail in Section 7. One specific aspect of such a sequential monitoring framework is determining which data should be used at a given analysis or look time. This becomes a larger issue when one is using exposure matching or stratification. Statistically to incorporate a sequential monitoring boundary the exposure matching and stratification categories at previous looks must remain the same (i.e. attempting to keep the population used in previous looks static). However, there are choices in which data to use to fit the current look's DRS and PS models to determine new disease/exposure matches or disease/exposure stratification categories. For this simulation study we have explored two approaches: lookwise estimation and cumulative estimation.

a. Lookwise Estimation

Lookwise estimation uses the current look's incident users to build PS and DRS models.[1] Specifically, let N_t and n_t represent the look specific population size (treatment of interest and comparator treatment) and treatment of interest population size (i.e. number of individuals who received the treatment of

interest (Z=1)) in interval t, respectively. In lookwise estimation, for look t, a propensity score model was fit and each individual's PS(t) was estimated using only the current look's population, N_t and n_t. Then matching and stratification were performed using the look specific PS from look t. An individual remained in his or her stratum in current and subsequent analyses of future looks.

b. Cumulative Estimation

Cumulative estimation uses the cumulative population to the current look (from beginning of surveillance) in building PS and DRS models. Let $\dot{N}(k) = \sum_{t=1}^{k} N_t$ and $\dot{n}(k) = \sum_{t=1}^{k} n_t$ represent the cumulative total and treatment of interest population sizes up to kth look, respectively. $PS^c(k)$ denotes the cumulative PS estimated with population sizes $\dot{N}(k)$ and $\dot{n}(k)$. In matching, comparators who were matched to the treatment of interest in previous intervals 1,2,…,(t-1) remained in their previous matched strata. Similarly for stratification those were classified as being within stratum s in previous looks remained in stratum s for all future looks. Thus, only the incident users in a current look were either matched or reclassified into strata.

Both lookwise and cumulative estimation methods used similar PS and DRS models for computing PSs and DRSs. The PS for an individual is the probability of receiving the treatment of interest given his/her covariate values. For incident users of the treatment of interest and comparators in each look, we fit a logistic regression model with the treatment status as the dependent variable and variables x_1,…, x_{10} as predictors (propensity score models). After all parameter coefficients were estimated in the propensity score model, PSs were calculated for each individual as

$$PS = \frac{\exp(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+\hat{\beta}_3 x_3 + \cdots +\hat{\beta}_{10} x_{10})}{1 + \exp(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+\hat{\beta}_3 x_3 + \cdots +\hat{\beta}_{10} x_{10})}$$

where $\hat{\beta}_0$, $\hat{\beta}_1$, …, $\hat{\beta}_{10}$ are estimated coefficients of intercept and confounders using standard maximum likelihood estimation.

The DRS for an individual is the probability of experiencing an outcome (adverse event) given his/her covariate values. In creating DRS model, we included all covariates (x_1, …, x_{10}) and the treatment variable in the model. [2,3,4] Thus, the DRS model was the same as model (2). Inclusion of the treatment in the DRS model is critical. Otherwise the effect of treatment on the outcome will not be represented in the model and the true association of confounders with the outcome will be distorted. Consequently the DRS will be wrong and the results from outcome model will be invalid. Our preliminary results showed that the treatment effect approached to the null if treatment status was not included in DRS model. After all parameters were estimated, DRS was calculated as

$$DRS = \frac{\exp(\hat{\theta}_0+\hat{\theta}_1 x_1+\hat{\theta}_2 x_2+\hat{\theta}_3 x_3 + \cdots +\hat{\theta}_{10} x_{10})}{1 + \exp(\hat{\theta}_0+\hat{\theta}_1 x_1+\hat{\theta}_2 x_2+\hat{\theta}_3 x_3 + \cdots +\hat{\theta}_{10} x_{10})}$$

where $\hat{\theta}_0$, $\hat{\theta}_1$,…, $\hat{\theta}_{10}$were estimated from DRS model. Note that even though treatment was included in the DRS model which was used to calculate the parameters, the term $\hat{\theta}_z Z$ was not included (i.e., Z=0) for the entire cohort, in the calculation of DRSs for individual participants.[3]

In creating PS or DRS models, we anticipated quasi-data separation would occur when treatment or outcome numbers are low (i.e., 20-50 outcome or treatments of interest). To solve this issue in a simulation setting, we used the backward selection approach with a p-value <0.5 as the inclusion criteria to select covariates. In this instance, it was not feasible to select variables based on a *priori* knowledge, as we were performing 1000 replicate simulations per each combination of simulation parameters.

Once DRS and PS were calculated, we performed score matching or stratification to create analytic datasets for sequential analyses as detailed in the following section.

4. Matching and Stratification

We performed 1:1 and 1:4 matching on closest values of PS or DRS for different treatment and outcome scenarios ranging from rare to common. Participants matched in previous looks were not re-matched in subsequent time intervals. As incident users were used, individuals receiving the treatment of interest who were not matched in previous looks were not re-used in subsequent looks. For most simulations, we limited the matching ratio to 1:4 because little additional statistical power is gained with additional matches. However, for rare treatment and rare outcome scenarios, we also performed 1:8 matching in order to increase the number of individuals and adverse events in sequential analyses.

We also stratified participants based on percentiles of PS or DRS and performed stratified analyses. At each look, PSs and DRSs were computed using either lookwise or cumulative estimation methods, and each subject was assigned to a stratum based on his/her scores; the subject remained in the stratum throughout following sequential analyses. We explored the impact of different number of strata on parameter estimates and found no significant difference among 5, 10, and 20 strata (Appendix Tables A2 – A3). Thus, we used 10 strata for all simulations.

5. Evaluating Treatment-Outcome Associations after Matching or Stratification

We used conditional logistic regression to evaluate treatment-outcome associations for all analyses matched on PSs or DRSs.

For stratified analyses, we explored different regression methods including Mantel-Haenszel logistic regression[5], which required pooling results from individual stratum, Firth method which corrects bias when outcome is rare in logistic regression[6], and fixed effects and random effects logistic regression in which stratum indicators were treated as fixed effects and random effects[7], respectively (Tables A2-A3). We obtained comparable estimates of treatment effects for conditional, fixed effects, and random effects regression and we opted to perform all stratified analyses using conditional logistic regression, as this method is commonly used in the existing literature. In addition, by using the same outcome model for matching and stratification, we could better ensure that any difference observed in the results could be attributed to different methodological approaches rather than analytic models.

6. Evaluating Treatment-Outcome Associations Using Regression

The regression model for evaluating the association between treatment and outcome was the same as model (2). When we experienced a model convergence problem for rare (~20 events) and moderately rare event (~50 events) scenarios we resorted to using a backward selection procedure with p<0.5 as the significance level for inclusion of covariates, as it was not possible to individually select appropriate confounders for each of the 1000 replicate simulations.

7. Methods for Longitudinal Sequential Estimation

To incorporate group sequential monitoring to form boundaries we used Lan-Demets Group Sequential approach.[8] This method uses an error spending approach and derives asymptotically normal sequential boundaries taking into account sequential testing of the data following a specified error spending function. Note that in Mini-Sentinel most boundaries applied have used exact methods to form sequential boundaries, but since the emphasis of this workgroup was to compare PS versus DRS we simplified the boundary formation to allow for more extensive simulations to be conducted towards the goals of the workgroup. We maintained an overall one-sided significance level of 0.05 in the sequential analyses. For this simulation we chose to use the Pocock error spending function, $\alpha(t) = \log\left[1 + (\exp(\mathrm{t}) - 1)\frac{\dot{N}_t}{\dot{N}_T}\right]\alpha$ (where α is the overall type I error), that attempts to spend the error so that on the standardized test statistic scale the boundary is approximately flat.[21] This is the most commonly used shape of the boundary for postmarket surveillance. The test statistic to compare to the sequential boundary was the standardized Wald test statistic, $\hat{\theta}_z/\sqrt{\mathrm{var}(\hat{\theta}_z)}$ and was calculated at each pre-specified look time (t=1,…,*10*). These boundaries are the critical values of the sequential hypothesis tests. After each interim test, if the test statistic is less than the boundary at time t, the trial is continued. When the test statistic is greater than the boundary at time t, the null hypothesis of equal means is rejected and the trial is stopped early.

8. Parameters for Simulating Treatment and Outcome

We simulated 1000 replicate populations, each consisting of 50,000 participants (5,000 per year). We created replicate cohorts by looping back through the code with different starting seed numbers in the randomization routines.

In each population, age was a normally distributed variable with a mean of 50 and a standard deviation of five. This resulted in an age range of 30 – 70 years. The population consisted of 70% of one gender (in this instance, assumed to be male) and 30% of the other (female). We assumed the prevalence of acute disease to be 5% during any given time interval and the prevalence of chronic disease to be 5%. We also included follow-up time and an interaction term between follow-up time and chronic disease to model an increase in chronic disease over time in simulating treatment status. In addition, we included six other covariates with a prevalence of 5%.

We simulated three different scenarios related to strengths of confounding effects. In the first scenario, the strengths of the associations between the confounders x_1,…, x_4 and the treatment and confounders, x_1,…, x_4 and the outcome were the same. In the second scenario, these confounders were more strongly associated with the treatment. In the third, they were more strongly associated with the outcome. Varying the confounding strengths enabled comparison between the PS and DRS methods of confounding adjustment in observational studies. Details of parameter coefficients for these three scenarios are provided in Table 1, for one example simulation (common treatment (~500 in the first year increasing to ~1000 in year 10) and rare outcome (~20 outcomes per year)).

Table 1. Parameter coefficients for different confounding strength

	CONFOUNDING STRENGTH SCENARIOS					
	1 (equal)		2 (toward treatment)		3 (toward outcome)	
Parameters	**Coefficients for the treatment probabilities**	**Coefficients for the outcome probabilities**	**Coefficients for the treatment probabilities**	**Coefficients for the outcome probabilities**	**Coefficients for the treatment probabilities**	**Coefficients for the outcome probabilities**
Intercept[+]	-2.5	-6.0	-3.3	-6.0	-2.5	-6.85
Age	0.002	0.002	0.01	0.002	0.002	0.01
Gender	0.2	0.2	1.0	0.2	0.2	1.0
Acute Disease	0.15	0.15	0.75	0.15	0.15	0.75
Chronic Disease	0.15	0.15	0.75	0.15	0.15	0.75
Time (t)	0.1	0.1	0.1	0.1	0.1	0.1
Chronic Disease*t	0.2	NA	0.2	NA	0.2	NA
x_5	0.1	0.1	0.1	0.1	0.1	0.1
x_6	0.1	0.1	0.1	0.1	0.1	0.1
x_7	0.1	0.1	0.1	0.1	0.1	0.1
x_8	0.1	0.1	0.1	0.1	0.1	0.1
x_9	0.1	0.1	0.1	0.1	0.1	0.1
x_{10}	0.1	0.1	0.1	0.1	0.1	0.1
treatment	NA	0, 0.69	NA	0, 0.69	NA	0, 0.69

[+]Intercept parameters in this table are based on the rare outcome (<20/5000) and common treatment (>500/5000) scenario. Intercept coefficients for all scenarios are available in Appendix Table A1.

To make valid comparisons of type I rates and empirical power across different scenarios, we chose different intercepts for different scenarios so that the frequencies of treatments and outcomes were similar. We varied the intercepts in models (1) and (2) to achieve different prevalence of treatment of interest and outcomes ranging from rare, moderately rare, and common (Table 2). The simulated rare treatment and rare outcome may be less applicable to typical Mini-Sentinel projects because the treatment and outcome numbers are very low. A sample of frequencies of treatments and outcomes are listed in Table 3a (rare outcome and common treatment) and Table 3b (common treatment and common outcome).

Table 2. Simulated treatment and outcome

TREATMENT ($\bar{n}$)		OUTCOME ($\bar{y}$)	
	Increasing treatment frequency look 1 to 10		Average number of outcomes (consistent all looks)
Rare	25	Rare	<20
Rare	25	Moderate	50
Rare	25	Common	200-250
Moderate	50	Rare	<20
Moderate	50	Moderate	50
Moderate	50	Common	200-250
Common	>500	Rare	<20

TREATMENT ($\bar{n}$)		OUTCOME ($\bar{y}$)	
Common	>500	Moderate	50
Common	>500	Common	200-250

Table 3a. Average number of outcomes ($\bar{y}$) and average number of participants receiving the treatment of interest ($\bar{n}$) across years in common treatment and rare outcome scenarios (≤20 outcomes; OR=1)

	CONFOUNDING STRENGTH SCENARIOS					
	SCENARIO 1		SCENARIO 2		SCENARIO 3	
YEAR	$\bar{y}$	$\bar{n}$	$\bar{y}$	$\bar{n}$	$\bar{y}$	$\bar{n}$
1	17	514	17	516	16	514
2	17	549	18	547	16	549
3	16	614	17	621	16	614
4	17	662	17	650	17	662
5	17	727	17	727	17	727
6	18	759	18	771	17	759
7	18	849	18	835	17	849
8	18	916	18	893	17	916
9	18	997	18	968	17	997
10	18	1080	18	1057	18	1080

Table 3b. Average number of outcomes ($\bar{y}$) and average number of participants receiving the treatment of interest ($\bar{n}$) across years in common treatment and outcome scenarios (>200 outcomes; OR=1)

	SCENARIO 1		SCENARIO 2		SCENARIO 3	
YEAR	$\bar{y}$	$\bar{n}$	$\bar{y}$	$\bar{n}$	$\bar{y}$	$\bar{n}$
1	218	508	218	513	213	507
2	218	553	218	556	212	553
3	217	606	217	606	213	606
4	218	656	219	656	214	656
5	217	718	217	714	213	718
6	219	779	219	773	214	779
7	219	848	219	838	214	849
8	218	919	218	906	212	919
9	217	1000	217	980	214	1000
10	218	1078	218	1054	214	1078

9. Evaluation Measures

We evaluated the different confounder adjustment methods according to the following evaluation metrics:

- type I error rates (false positive rates) which were calculated as the percentage of datasets that rejected the null hypothesis $\theta_Z = 0$ when data were simulated under $\theta_Z = 0$;
- bias which was calculated as the difference between the estimated treatment parameter θ_Z and the true treatment parameter under which the data were simulated;
- power to detect an association which was calculated as the percentage of datasets that rejected the null hypothesis $\theta_Z = 0$ when simulations were performed under the true alternative $\theta_Z > 0$; and
- time to signal detection which is from the start of monitoring (i.e., year one in simulation) to the time when the test statistic is greater than the boundary.

C. RESULTS

The results from all simulations performed on a yearly basis are presented in Appendix B. Our main observations in regard to type I error rates, empirical power, treatment effect estimates, and time to signal for the different analysis methods are summarized below.

1. In general, we observed no difference between lookwise and cumulative estimation methods.
2. In general, we observed no difference between PS and DRS approaches. This result is consistent with a cross-sectional observational study by Arbogast and Ray[3], showing that DRS and PS approaches yielded comparable risk estimates when DRS and PS models were correctly specified.
3. Empirical power and type I error rates did not differ with different strengths of associations 1) between confounders and treatment, and 2) between confounders and outcome, except in the following scenarios:

    a) Matching 1:4 or stratifying on PSs or DRSs in rare treatment and rare outcome settings: When treatment and outcome counts were rare, the type I error rates were higher for scenario 1 (equal association strength of confounders with treatment of interest and outcome) when matching 1:4 on PSs or DRSs compared to scenarios 2 (stronger association of confounders with treatment of interest than outcome) and 3 (stronger association of confounders with outcome than treatment of interest) (Table 4). In addition, 1:4 matching on PSs yielded a higher empirical power (14.7%) for scenario 1 compared to scenarios 2 and 3 (5.1% and 5.8%) (Table 5). Similar results were observed in regard to power when matching 1:4 on DRSs;

    b) Matching 1:4 on PSs and DRSs in the common treatment and common outcome setting. When matching 1:4 on PS, there was a 96% increase in type I error rates between scenarios 1 and 2 (5.6% vs 111%), and a 69% decrease between scenarios 2 and 3 (11% vs 3.4). Similar results were observed when matching 1:4 and stratifying on DRSs.

Table 4. Type I error rates with lookwise estimation method under OR=1 ($\theta_Z = 0$)

METHOD	RARE TREATMENT AND RARE OUTCOME			MODERATE TREATMENT AND MODERATE OUTCOME			COMMON TREATMENT AND COMMON OUTCOME			RARE TREATMENT AND MODERATE OUTCOME			MODERATE TREATMENT AND RARE OUTCOME		
	Scenario			Scenario			Scenario			Scenario			Scenario		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
PS 1:1	0.0	0	0	0.7	0.7	0.2	5.8	5.5	2.2	0.0	0	0	0	0	0
PS 1:4	2.0	0.9	1.0	5.8	5.9	4.5	5.6	11.0	3.4	3.8	4.3	3.9	2.8	3.4	3.7
PS stratification	10.4	11	10.9	8.6	9.7	8.5	5.5	5.0	2.9	8.7	8.7	10.5	9.0	8.6	9.0
DRS 1:1	0.0	0	0	0.5	0.8	0.2	6.3	11.4	3.7	0.0	0	0	0	0	0
DRS 1:4	1.6	0.8	0.5	5.2	6.8	4.5	6.3	17.8	5.3	4.6	4.1	3.8	3.1	3.5	3.8
DRS stratification	10.1	10.7	11.6	8.8	9.9	8.4	6.8	18.1	4.4	9.0	9.3	10.1	9.2	8.6	8.1

Table 5. Empirical power with lookwise estimation method for computing PS and DRS under OR=2 ($\theta_Z = 0.693$)

METHOD	RARE TREATMENT AND RARE OUTCOME			MODERATE TREATMENT AND MODERATE OUTCOME			COMMON TREATMENT AND COMMON OUTCOME			RARE TREATMENT AND MODERATE OUTCOME			MODERATE TREATMENT AND RARE OUTCOME		
	Scenario			Scenario			Scenario			Scenario			Scenario		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
PS 1:1	0.0	0.0	0.0	19.7	16.9	15.8	100	100	100	1.0	2.1	1.4	0.7	1.5	0.8
PS 1:4	14.7	5.1	5.8	50.9	49.4	50.0	100	100	100	28.2	28.8	30.3	24.8	30.6	28.1
PS strati-fication	36.4	30.6	30.8	65.7	66.8	65.4	100	100	100	44.8	47.4	48.5	39.7	44.4	46.6
DRS 1:1	0.0	0.0	0.0	17.2	16.1	16.1	100	100	100	1.3	2.0	1.5	0.6	1.5	0.7
DRS 1:4	13.3	4.7	5.7	49.6	50.9	50.9	100	100	100	30.9	30.4	30.8	24.3	29.7	27.3
DRS stra-tification	36.0	30.6	31.0	65.9	67.5	65.2	100	100	100	45.2	49.4	49.1	38.9	44.2	45.1

4. Matching methods using a 1:1 ratio yielded lower empirical power: When both treatment and outcomes were rare or moderately rare, we observed very different results in empirical power for 1:1 matching compared to 1:4 matching (Table 6). The 1:4 matching had clear benefits in empirical power when outcome or treatment counts were relatively low, but the benefits were less apparent when treatments or outcomes were common. The ratio of the empirical power between 1:4 matching versus 1:1 matching decreased when there were more treatments and/or outcomes. For example, when the treatment of interest and outcome were rare, the ratio was infinite (14.7% empirical power for 1:4 matching versus zero empirical power for 1:1); when the outcome became moderate, the ratio was 28.2 (28.2% empirical power for 1:4 matching versus 1% empirical power for 1:1); when the outcome became common, the ratio dropped to 1.72 (72.6% empirical power for 1:4 matching versus 42.0% empirical power for 1:1).

In the rare outcome setting, the ratio was infinite (14.7% empirical power for 1:4 matching versus zero empirical power for 1:1 matching) when treatment was rare; it decreased to 35.4 (24.8% empirical power for 1:4 matching versus 0.7% empirical power for 1:1 matching) when treatment was moderately rare; it further decreased to 1.3 (95.5% empirical power for 1:4 matching versus 76.4% empirical power for 1:1 matching) when treatment was common. When the OR increased to 5, we observed similar trends (Appendix Table A6a, A6b, and A6c).
We also explored 1:8 matching on PS using the lookwise estimation method for scenario 1 (equal strengths of associations between confounders and the treatment and confounders and the outcome). This increased empirical power to 22.8%, which is significantly higher than the 14.7% observed for 1:4 matching. However, it remained lower than the 36.4% we observed when stratifying on PS.

Table 6. Empirical power comparing 1:1 and 1:4 matching with lookwise estimation method and scenario 1 of confounding strength under OR=2 ($\boldsymbol{\theta_Z = 0.693}$)

METHOD	RARE TREATMENT			MODERATE TREATMENT			COMMON TREATMENT		
	Rare outcome	Moderate outcome	Common outcome	Rare outcome	Moderate outcome	Common outcome	Rare outcome	Moderate outcome	Common outcome
PS 1:1	0.0	1.0	42.0	0.7	19.7	77.4	76.4	99.8	100
PS 1:4	14.7	28.2	72.6	24.8	50.9	93.4	95.5	100	100
DRS 1:1	0.0	1.3	44.2	0.6	17.2	78.0	78.8	100	100
DRS 1:4	13.3	30.9	73.3	24.3	49.6	94.0	95.7	100	100

5. Stratification: yielded higher type I error rates and inflated empirical power in all scenarios except when the treatment of interest was common. This is likely caused by overestimated treatment effects as discussed below.

   Stratification kept all adverse events BUT resulted in higher than nominal type I error rates (2.0% for PS 1:4 matching versus 10.4% for PS stratification) when treatment and outcome were rare, indicating that using this approach may result in inflated power (Table 7). Table 8 shows the mean treatment effect coefficients and standard deviations for all the approaches by years of follow-up, and for those datasets that had valid treatment effect estimates. A treatment effect estimate between -12 and 12 in SAS output was considered valid. Examination of 1000 treatment effect estimates confirmed the need to exclude extreme values. In general, for the matching approach, the 1:1 matching ratio underestimated treatment effects. For example, in year 1, the mean parameter estimates for records with valid estimates were 0.001 and 0.002 for PS and DRS approaches, respectively. These values were well below the simulated 0.693. These estimates increased over the years, but only reached to a maximum value of 0.402 for PS 1:1 matching approach and 0.432 for DRS 1:1 matching approach at year 10. This pattern explains the lower empirical power of 1:1 matching when treatment and outcomes are rare; Estimates from 1:4 matching had higher initial values (0.08-0.09) which increased from year 1 to year 5 to reach levels exceeding the simulated 0.693 (~0.90) and decreased afterward to levels closer to the true value of the treatment parameter by year 10 (0.69-0.70). At year 1, the valid initial estimates from stratification (N~200) were close to or exceeded the simulated level (PS: 0.89; DRS: 0.65). These estimates then increased from year 1 to year 3 and decreased afterwards to levels slightly below the true value (PS: 0.61; DRS: 0.60). Compared to 1:4 matching,

stratification overestimated treatment effects in the early years (2, 3, 4 and 5). This may explain the inflated empirical power observed when using stratification in the annual sequential analyses.

The inflation of empirical power by stratification was greatest when both treatment and outcome were rare, and lessened with increasing incidence of treatment and/or outcomes (Table 9). For example, in rare treatment and rare outcome settings, the ratio of empirical power between stratification and 1:4 matching on PS was 2.48 (36.4%/14.7%=2.48). This decreased to 1.59 (44.8%/28.2%=1.59) in scenarios in which the outcome was moderately rare, and further decreased to 1.14 (82.7%/72.6%=1.14) in common outcome scenarios. With the increased incidence of treatment, we also observed a decrease in the inflation of empirical power. For example, the ratio of powers between stratification and PS matching1:4 was 2.48 (36.4%/14.7%=2.48) when both treatment and outcome were rare. This ratio decreased to 1.60 in settings with moderately rare treatment, and further decreased to 1.01 in scenarios where treatment was common (96.2%/95.5%=1.01). Similar patterns were observed when matching and stratifying on DRSs, and for scenarios 2 and 3.

Table 7. Rare treatment rare outcome: regression and lookwise estimation method and scenario 1 of confounding strength

		TREATMENT PARAMETER $\theta_Z = 0.693$ (OR=2)				
Design and Method	$\theta_Z = 0$ **Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0.0	0.0	NA	0.40 (0.71), 712	NA	NA
PS Matching 1:4	2.0	14.7	1.94 (0.48)	0.74 (0.80), 929	8.19 (1.68)	8 (5 10)
PS Stratification	10.4	36.4	1.91 (0.66)	0.93 (0.96), 932	4.95 (3.15)	5 (1 10)
DRS Matching 1:1	0.0	0.0	NA	0.44 (0.75), 711	NA	NA
DRS Matching 1:4	1.6	13.3	1.94 (0.45)	0.73 (0.78), 930	8.11 (1.70)	8 (5 10)
DRS Stratification	10.1	36.0	1.93 (0.66)	0.93 (0.96), 929	4.84 (3.13)	4 (1 10)
Regression	10	35.9	1.93 (0.67)	1.50 (0.78), 572	4.81 (3.08)	4 (1 10)

Table 8. Number of valid treatment effects (N) and treatment effects (standard deviation) when treatment and outcome were rare for scenario 1 of confounding strength under θ_Z=0.693

YEAR	PS						DRS			
	1:1 match		1:4 match		stratification		1:1 match		1:4 match	
	N	Mean $\widehat{\theta}_Z$ (std)	N	Mean $\widehat{\theta}_Z$ (std)	N	Mean $\widehat{\theta}_Z$ (std)	N	Mean $\widehat{\theta}_Z$ (std)	N	Mean $\widehat{\theta}_Z$ (std)
1	798	0.001 (0.052)	646	0.085 (0.352)	205	-0.893 (6.001)	776	0.002 (0.047)	622	0.077 (0.326)
2	643	-0.0002 (0.103)	491	0.426 (0.664)	300	1.767 (1.111)	635	0.001 (0.101)	494	0.388 (0.637)
3	561	0.020 (0.208)	475	0.725 (0.711)	420	1.449 (0.378)	554	0.005 (0.191)	476	0.707 (0.723)
4	514	0.054 (0.332)	520	0.878 (0.714)	519	1.199 (0.436)	517	0.029 (0.314)	511	0.867 (0.737)
5	488	0.101 (0.446)	589	0.932 (0.719)	625	1.017 (0.480)	497	0.089 (0.445)	605	0.887 (0.741)
6	515	0.146 (0.523)	676	0.902 (0.719)	703	0.882 (0.505)	502	0.165 (0.538)	678	0.898 (0.744)
7	527	0.209 (0.597)	765	0.880 (0.757)	781	0.781 (0.526)	550	0.242 (0.616)	770	0.862 (0.750)
8	583	0.266 (0.667)	841	0.782 (0.751)	849	0.676 (0.553)	595	0.311 (0.681)	841	0.779 (0.769)
9	644	0.372 (0.696)	893	0.747 (0.760)	898	0.640 (0.558)	664	0.375 (0.711)	896	0.741 (0.744)
10	712	0.402 (0.712)	929	0.703 (0.748)	932	0.605 (0.559)	711	0.432 (0.745)	930	0.685 (0.720)

Table 8 continued

YEAR	DRS		REGRESSION	
	STRATIFICATION			
	N	Mean $\widehat{\theta}_Z$ (std)	N	Mean $\widehat{\theta}_Z$ (std)
1	200	-0.654 (5.847)	152	2.632 (0.448)
2	300	1.764 (1.112)	298	1.839 (0.371)
3	420	1.448 (0.374)	420	1.442 (0.372)
4	518	1.199 (0.429)	490	1.224 (0.420)
5	622	1.019 (0.478)	401	1.268 (0.405)
6	700	0.883 (0.503)	385	1.274 (0.314)
7	778	0.778 (0.525)	483	1.132 (0.316)
8	846	0.674 (0.554)	558	1.011 (0.335)
9	895	0.637 (0.559)	579	0.972 (0.327)
10	929	0.600 (0.560)	550	0.982 (0.291)

Table 9. Stratification approach inflated empirical power (%) and the problem is lessened with increasing number of treatments and number of outcomes for scenario 1 of confounding strength under θ_z=0.693

METHOD	RARE TREATMENT			MODERATE TREATMENT			COMMON TREATMENT		
	Rare outcome	Moderate outcome	Common outcome	Rare outcome	Moderate outcome	Common outcome	Rare outcome	Moderate outcome	Common outcome
PS 1:4	14.7	28.2	72.6	24.8	50.9	93.4	95.5	100	100
PS strat-ification	36.4	44.8	82.7	39.7	65.7	97.6	96.2	100	100
DRS 1:4	13.3	30.9	73.3	24.3	49.6	94.0	95.7	100	100
DRS stra-tification	36.0	45.2	83.7	38.9	65.9	97.0	95.8	100	100
Regress-ion	35.9	44.9	82.8	39.2	65.1	96.4	96.5	100	100

6. Regression: Similar to stratification, regression yielded higher type I error rates and inflated empirical power in all scenarios except when the treatment of interest was common. This is likely caused by overestimated treatment effects (Table 7-9).

7. Comparison of annual and biennial sequential analyses: To examine whether less frequent sequential analyses would increase empirical power, we conducted simulations on a biennial basis, where we created scores and performed sequential analyses every two years. The upper boundaries using Lan-Demets Group Sequential approach for rejecting the null hypothesis are given in Table 10 for both the primary annual analyses and the biennial analyses. Because there were fewer looks during the ten years of follow-up, the upper boundaries for the biennial sequential analyses were lower than those for annual sequential analyses. Biennial sequential analyses increased the empirical power for matching approaches, although not substantially, and the empirical power for stratification remained stable or decreased (Table 11). For example, in scenario 1 (equal confounding strength between confounders and treatment and confounders and the outcome), the empirical power increased from 14.7% to 17.6% when biennial sequential was conducted for 1:4 matching approach whereas the empirical power decreased slightly for the PS stratification approach. Consistent with the power changes, the treatment effects estimates increased for the 1:4 matching approach at years 4 and 6 (Table 12) while the treatment effects estimates did not change for the stratification approach. Similar results were observed for scenarios 2 (stronger associations between confounders and treatment) and 3 (stronger associations between confounders and outcome). The increase in empirical power was less when treatment and outcome were moderately rare (data not shown).

Table 10. Boundaries based on Lan-Demets Group Sequential approach

LOOK	ANNUAL	BIENNIAL	LOOK	ANNUAL	BIENNIAL
1	2.412202		6	2.225038	2.113224
2	2.362529	2.176211	7	2.204309	
3	2.316565		8	2.186592	2.089523
4	2.279599	2.143710	9	2.171227	
5	2.249699		10	2.157767	2.070907

Table 11. Comparison of annual and biennial sequential analyses for rare treatment rare outcome using lookwise estimation method for creating PS or DRS scores for confounding strength scenarios 1 , 2 and 3 under θ_Z=0.693

		ANNUAL SEQUENTIAL ANALYSES			BIENNIAL SEQUENTIAL ANALYSES		
CONFO-UNDING SCENAR-IOS	**Design and Method**	**Signaled %**	**Mean $\widehat{\theta}_Z$ (std), Signaled and Not, N**	**Mean Time to Signal Detection (STD)**	**Signaled %**	**Mean $\widehat{\theta}_Z$ (std), Signaled and Not, N**	**Mean Time to Signal Detection (STD)**
1	PS Matching 1:1	0.0	0.40 (0.71), 712	NA	0.1	0.41 (0.73)	10 (NA)
	PS Matching 1:4	14.7	0.74 (0.80), 929	8.19 (1.68)	17.6	0.79 (0.87), 931	7.83 (1.92)
	Stratification	36.4	0.93 (0.96), 932	4.95 (3.15)	32.9	0.80 (0.78), 932	5.98 (2.82)
	DRS Matching 1:1	0.0	0.44 (0.75), 711	NA	0.1	0.40 (0.71), 724	10 (NA)
	DRS Matching 1:4	13.3	0.73 (0.78), 930	8.11 (1.70)	16.1	0.78 (0.83), 923	7.90 (2.06)
	Stratification	36.0	0.93 (0.96), 929	4.84 (3.13)	32.7	0.79 (0.78), 931	5.96 (2.80)
2	PS Matching 1:1	0.0	0.23 (0.63), 711	NA	0.0	0.25 (0.64), 568	NA
	PS Matching 1:4	5.1	0.85 (0.77), 799	8.67 (1.70)	7.6	0.83 (0.81), 798	8.26 (2.14)
	Stratification	30.6	1.14 (1.01), 818	4.86 (3.29)	31.8	1.04 (0.85), 834	5.58 (3.17)
	DRS Matching 1:1	0.0	0.22 (0.62), 569	NA	0.0	0.28 (0.61), 588	NA
	DRS Matching 1:4	4.7	0.84 (0.79), 805	8.70 (1.63)	6.8	0.87 (0.77), 814	8.16 (2.07)
	Stratification	30.6	1.14 (1.02), 818	4.82 (3.32)	30.8	1.02 (0.84), 834	5.64 (3.07)
3	PS Matching 1:1	0.0	0.19 (0.60), 528	NA	0.0	0.25 (0.56), 504	NA
	PS Matching 1:4	5.8	0.92 (0.81), 762	8.52 (1.50)	8.8	0.94 (0.81), 758	8.73 (1.56)
	Stratification	30.8	1.26 (1.06), 784	4.45 (3.19)	34.8	1.11 (1.07), 796	5.02 (3.05)
	DRS Matching 1:1	0.0	0.22 (0.57), 526	NA	0.0	0.16 (0.57), 490	NA
	DRS Matching 1:4	5.7	0.92 (0.80), 755	8.42 (1.69)	8.2	0.98 (0.80), 760	8.49 (1.72)
	Stratification	31.0	1.27 (1.05), 783	4.40 (3.17)	30.0	1.10 (1.07), 776	5.11 (3.05)

Table 12. Comparison of annual and biennial sequential analyses for rare treatment rare outcome using lookwise estimation method for scenario 1: treatment effect estimate over time under θ_Z=0.693

YEAR	1:4 MATCHING ANNUAL		1:4 MATCHING BIANNUAL		STRATIFICATION ANNUAL		STRATIFICATION BIANNUAL	
	N*	Mean $\widehat{\theta}_Z$ (std)	N	Mean $\widehat{\theta}_Z$ (std)	N	Mean $\widehat{\theta}_Z$ (std)	N	Mean $\widehat{\theta}_Z$ (std)
2	491	0.426 (0.664)	482	0.411 (0.670)	301	1.722 (1.331)	300	1.767 (1.111)
4	520	0.878 (0.714)	515	0.914 (0.710)	519	1.200 (0.434)	519	1.199 (0.436)
6	676	0.902 (0.719)	678	0.944 (0.745)	703	0.885 (0.501)	703	0.882 (0.505)
8	841	0.782 (0.751)	832	0.803 (0.803)	849	0.679 (0.551)	849	0.676 (0.553)
10	929	0.703 (0.748)	931	0.715 (0.773)	932	0.607 (0.558)	932	0.605 (0.559)

*Number of valid estimates of treatment effect out of 1000 replicates.

III. METHODOLOGICAL EXAMPLE

A. SELECTION CRITERIA

Our criteria for selection of a methodological example included the following: the selected medical product should have known adverse events and have known confounders from prior research; it should have varying use in the population during the given study timeframe, and should be prescribed to participants with varying characteristics over time; the outcomes of interest need to range from rare to more common and should have available data in the MSDD during the study timeframe. As examples from previous or ongoing M-S projects did not meet these criteria, we selected Cyclo-Oxygenase-2 (COX-2) Inhibitors, a class of Non-Steroid Anti-Inflammatory Drugs (NSAIDs), as our exposure of interest and Non-Selective NSAIDs as the active comparator, and evaluated the association between these NSAIDs and gastrointestinal (GI) bleeding events.

B. METHODS

1. NSAID Use and Risk of Adverse Events

NSAIDS are frequently prescribed to alleviate pain due to inflammation caused by ailments such as osteoarthritis. Currently available NSAIDs operate by blocking COX enzymes. These enzymes include COX-1 which protects the lining of the stomach from acid, COX-2, found in joints and muscle, which mediates effects on pain and inflammation, and COX-3, located in the cerebral cortex, which is possibly associated with centrally mediated pain. Non-Selective NSAIDs that block COX-1 and COX-2 reduce pain, but may also cause gastrointestinal bleeding. On the other hand, COX-2 Selective Inhibitor drugs selectively block the COX-2 enzyme and were developed with the expectation that these should be safer with regard to gastrointestinal bleeding. Research studies have generally confirmed lower GI bleeding rates for COX-2 inhibitor use, but some studies have found them to be associated with increased risk of serious cardiovascular and other adverse events. [9,10]

The data for this example were extracted from selected Data Partners contributing to the MSDD. Two of the COX-2 Inhibitors were withdrawn from the market during this period (Table 13). In addition, research conducted between January 2002 and August 2004 indicated an increased risk of cardiovascular disease events with certain COX-2 inhibitors.

Table 13. Dates of FDA approval and alert or withdrawal for COX-2 Inhibitors

COX-2 INHIBITOR	APPROVAL DATE	DATE OF WITHDRAWAL OR RELEASE OF ALERT
celecoxib (brand name: Celebrex)	12/31/1998	March 2005,FDA issued an alert that it may be associated with increased CVD risk
valdecoxib (brand name: Vioxx)	5/21/199	September 2004, withdrawn from market
valdecoxib (brand name: Bextra)	11/19/2001	April 2005, withdrawn from market

2. Comparisons of Interest

a. Drug and Outcome Comparisons

In the primary analyses, we examined the association between Cox-2 Inhibitors and GI bleeding events using non-selective NSAIDs as an active comparator.

b. Methodological Comparisons

As in the simulation, we contrasted several analytic methods:

1. Propensity score and Disease Risk score adjustments
2. Matched and stratified adjustments
3. Two different methods of estimating the propensity and disease risk scores: 1) lookwise which used data of the current year in building score models; 2) cumulative which used data from times up to and including the current year. Although prior data is used to estimate new scores in the cumulative method, newly calculated scores are used only for the new users in a time interval for matching and stratification. Individuals who started their drugs in prior time intervals retain their original scores and matching or stratification selections.

Site specific propensity scores and disease risk scores were estimated and participants were matched or placed into strata within site. Outcome models used survival analysis and included the matching or stratification indicators as strata. Indicators for site were also included in the strata statement.

3. Data Source and Study Time Frame

We requested claims data recorded from January 2001 until December 2011 at four Data Partner sites that contributed to the MSDD: Aetna, Humana, and Kaiser Permanente in Northern California (KPNC) and Kaiser Permanente in Colorado (KPCO). Returned records showed that data were primarily available between 2008-2011 at two of the largest sites (Aetna and Humana). Thus, results in this report are restricted to that time frame. In addition, use of COX-2 inhibitors was uncommon at the smallest site

(KPCO) during this time frame and was insufficient to support site specific score estimates. Therefore, this site is not included in the primary results reported here.

4. Identification of Incident Users

We identified incident users of NSAIDs aged 18 years and older, among those who were health plan members between January 2008 and December 2011. All participants had to be enrolled in a health plan with medical and pharmacy benefits for a continuous period of six months (183 days) prior to the enrollment date (the washout period), and had no evidence of prior NSAID use. In addition, individuals with prior GI bleeding were excluded from analyses examining incident outcomes. As gaps in enrollment, pharmacy or medical benefits of 45 days or less usually represent administrative gaps rather than actual disenrollment, we ignored such gaps during participant recruitment.

We used National Drug Codes (NDCs) recorded in MSDD's outpatient pharmacy dispensing file to ascertain NSAID use. Thus, over-the-counter use was not captured. The NDCs used to identify users of COX-2 Inhibitors and Non-Selective NSAIDs in the analysis were obtained from the FirstDataBank.

5. Identification of Outcomes of Interest

Our primary outcome of interest was upper GI bleeding, including hospitalization for upper GI bleeding and Peptic Ulcer Disease. These diagnoses were identified using International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) codes 531.x, 532.x, 533.x, 534.x, and 578.x[9, 10, 11] (Table 14). The positive predictive value (PPV) for this composite algorithm is approximately 90%.[9, 10,12,13]

Table 14. ICD-9 codes and sources for outcome of interest

OUTCOME	DEFINITION NOTES	ICD-9-CM Codes	SOURCE/NOTES
GI BLEEDING	ICD-9 discharge diagnoses for hospitalization for GI bleeding, Peptic Ulcer Disease	531.x, 532.x, 533.x, 534.x, 535.x, 578.X	Diagnosis table and encounter table (to select inpatient events)

6. Follow-up

Analyses focused on outcomes occurring within 6 months of drug initiation. This relatively short time window was chosen for two reasons: 1) early adverse events would more likely reflect potential drug associations and 2) short-term drug use was common in this population (e.g. median time on drugs was 30 days). We followed participants from the index date until they experienced one of the following events: 1) occurrence of the outcome of interest, 2) completion of 183 days of follow-up, 3) death, 4) cessation of the use of the drug of interest or initiation of an alternative NSAID, 5) reaching the end of the study, i.e. December 31st 2011, or 6) disenrollment from the health plan.

7. Potential Confounders

Table 15 provides a list of potential confounders that we included in our analyses. We ascertained age at index date and sex from MSDD's demographic file, and current and past medical conditions during the washout period from ICD-9 diagnosis codes recorded during inpatient, outpatient, or emergency department visits from MSDD's diagnosis file.

Table 15. Covariates included in the PS, DRS, and regression models

CONFOUNDER	SOURCE/ICD-9-CM CODES
Calendar year	
Demographic Variables	
Age	MSDD demographic file
Gender	MSDD demographic file
Current or past medical conditions	
History of prior GI bleed or peptic ulcer disease	MSDD diagnosis file; ICD9 Codes: 530.82, 531.x, 532.x, 533.x, 534.x, 535.x, 578.x
Modified Charlson Comorbidity Score[14]	MSDD diagnosis file; see reference
Rheumatoid Arthritis	MSDD diagnosis file; ICD9 Codes:714
Osteoarthritis	MSDD diagnosis file; ICD9 Codes:715
Hypertension	MSDD diagnosis file; ICD9 Codes:401-405
History of MI	MSDD diagnosis file; ICD9 Codes:410
History of other CAD	MSDD diagnosis file: ICD9 Codes: 411-414
Diabetes	MSDD diagnosis file; ICD9 Codes:250
History of Ischemic stroke[11]	MSDD diagnosis file; ICD9 Codes:433,434, 436
History of non-traumatic intracerebral hemorrhage	MSDD diagnosis file; ICD9 Codes:430 to 432
Renal disease[14]	MSDD diagnosis file: ICD9 Codes: 403.01, 403.11,403.91, 404.02, 404.03, 404.12, 404.13, 404.92, 404.93, 582.x, 583.0-583.7, 585.x, 586.x, 588.0, V42.0, V45.1, V56.x
Use of Proton Pump inhibitors or H2 blockers	National drug codes obtained from the FirstDataBank
Use of anti-coagulant drugs	National drug codes obtained from the FirstDataBank

C. RESULTS

The primary analyses examining the association between COX-2 NSAID use versus non-selective NSAID comparators and GI bleed outcomes included n=2,688,965 patients with new dispensings. Table 16 shows the numbers of GI bleed outcomes within six months and the distribution of NSAID dispensings by year. GI Bleed outcomes within 6 months were rare, occurring in less than 0.2% of patients. COX-2 inhibitors accounted for less than 4% of the new NSAID dispensings in each year and use decreased slightly over the four years studied. Appendix C, Table C1 shows this same data for each of the three sites.

Table 16. Incident GI bleed outcome numbers and COX-2 exposure numbers by year (Total N: 2,688,965 excludes persons with history of GI bleed at time of NSAID initiation)

YEAR OF DRUG INITIATION	GI BLEED WITHIN 6 MONTHS		TYPE OF NSAID DRUG	
	No N (%)	Yes N (%)	Non-selective NSAID N (%)	COX-2 NSAID N (%)
2008	630714 (99.85%)	929 (0.15%)	606495 (96.02%)	25148 (3.98%)
2009	800308 (99.86%)	1143 (0.14%)	776778 (96.92%)	24673 (3.08%)
2010	660395 (99.87%)	889 (0.13%)	643921 (97.37%)	17363 (2.63%)
2011	598780 (99.86%)	807 (0.14%)	581236 (97.75%)	13351 (2.25%)

1: 4 Matched analyses (one COX-2 inhibitor matched to up to four nonselective NSAIDs)

We start by presenting results for matching based on the findings of the simulations which showed more consistent results with this method. As in the simulation, we created two matched cohorts, one with a fixed 1:1 ratio and the second with a fixed 4:1 ratio. COX-2 inhibitor use was rare and large numbers of nonselective NSAIDs were available as potential matches. Despite differing propensity score distributions, only 3 out of 80,535 COX-2 users did not have at least one match within a 0.05 caliper range and 94% had at least 4 matches within this caliper range. Table 17 presents results for both lookwise and cumulative estimated propensity score matches within ± 0.05 caliper. Results for the full cohort without this restriction are provided in Appendix C, Table 2C and those results trended slightly to the null as would be expected. In Table 17, results for the propensity scores show hazard ratios for nonselective NSAIDs versus COX-2 and the estimates are above one as would be expected given COX-2 is the referent category and these drugs are known to be protective against GI bleeding. The sequential test statistic was estimated as the parameter estimate divided by its standard error and the upper boundaries for signaling were calculated using the Lan-Demets Group Sequential approach based on the distribution of GI bleeding outcomes across the four one year time interval looks. Propensity scores estimated using lookwise versus cumulative methods showed no substantive differences and both signaled at the third look.

Table 17. Incident GI bleed outcome: results for Propensity Score matched samples (4:1) (4 Nonselective NSAIDs matched to 1 COX-2); Restricted to matches within 0.05

Look	Time[1]	Nonselective NSAID		COX-2 NSAID		HR[4] Nonselective NSAID vs COX-2	Sequential analysis test statistic[5]	Test statistic boundary[6]	Signal[7]
		GI bleeds$_{cum}$[2]	Person years$_{cum}$[3]	GI bleeds$_{cum}$	Person years$_{cum}$				
Lookwise PS[8]									
1	2008	264	44472	63	11510	1.17	1.119	2.1040	No
2	2009	566	88529	130	22824	1.21	1.952	2.0323	No
3	2010	769	119299	174	30701	1.23	2.391	2.0514	Yes
4	2011	932	143658	224	36877	1.15	1.813	2.0508	
Cumulative PS[9]									
1	2008	271	44480	63	11510	1.22	1.412	2.1040	No
2	2009	560	88578	130	22824	1.20	1.826	2.0323	No
3	2010	750	119480	174	30702	1.19	2.071	2.0514	Yes
4	2011	918	143957	224	36878	1.12	1.456	2.0508	

[1]Timing of each look; Later looks include information from the preceding time intervals

[2]GI bleeds$_{cum}$ is the total number of incident GI bleed events up to and including the year of each look (within each selected drug group)

[3]Person years$_{cum}$ is the total accumulated exposure time up to and including the year of each look (within each selected drug group)

[4]HR is the hazard ratio for nonselective NSAID vs COX-2 drugs for all times up to and including each look

[5]Test statistic is the drug comparison parameter estimate / standard error from the proportional hazards model

[6]Test statistic boundary: sequential analysis boundary estimate

[7]Signal indicates when the test statistic first exceeds the test statistic boundary

[8]Lookwise estimation used single year data when estimating scores

[9]Cumulative estimation used all years up to and including the current year when estimating scores

Table 18 presents similar results using the disease risk scores with COX-2 users matched up to four nonselective NSAID users. Similar to the propensity score matched analyses, matches were restricted to a caliper of 0.05 of the score although there were no users of COX-2 inhibitors who didn't have at least one match within this range and 99% had at least four. Appendix C, Table C3 has results with no restrictions and results were comparable though slightly moved towards the null. In Table 18, disease risk scores whether estimated by lookwise or cumulative methods both signaled at the second look.

Table 18. Incident GI bleed outcome: results for Disease Risk Score matched samples (4:1) (four Nonselective NSAIDs matched to one COX-2); Restricted to matches within 0.05

Look	Time[1]	Nonselective NSAID		COX-2 NSAID		HR[4] Nonselective NSAID vs COX-2	Sequential analysis test statistic[5]	Test statistic boundary[6]	Signal[7]
		GI bleeds$_{cum}$[2]	Person years$_{cum}$[3]	GI bleeds$_{cum}$	Person years$_{cum}$				
Lookwise DRS[8]									
1	2008	316	45320	63	11403	1.32	1.981	2.104	No
2	2009	663	89791	130	22602	1.32	2.848	2.0323	Yes
3	2010	889	120778	174	30366	1.33	3.376	2.0514	
4	2011	1052	144848	220	36388	1.23	2.817	2.0508	
Cumulative DRS[9]									
1	2008	318	45735	63	11510	1.29	1.820	2.104	No
2	2009	666	90663	130	22824	1.31	2.759	2.0323	Yes
3	2010	892	122089	174	30703	1.27	2.821	2.0514	
4	2011	1060	146772	224	36878	1.18	2.194	2.0508	

[1]Timing of each look; Later looks include information from the preceding time intervals
[2]GI bleeds$_{cum}$ is the total number of incident GI bleed events up to and including the year of each look (within each selected drug group) , only two sites included due to DRS data issues at third site
[3]Person years$_{cum}$ is the total accumulated exposure time up to and including the year of each look (within each selected drug group), only two sites included due to DRS data issues at third site
[4]HR is the hazard ratio for nonselective NSAID vs COX-2 drugs for all times up to and including each look
[5]Test Statistic is the drug comparison parameter estimate / standard error from the proportional hazards model
[6]Test statistic boundary: sequential analysis boundary estimate
[7]Signal indicates when the test statistic first exceeds the test statistic boundary
[8]Lookwise estimation used single year data when estimating scores
[9]Cumulative estimation used all years up to and including the current year when estimating scores

1: 1 Matched analyses (one COX-2 inhibitor matched to up to four nonselective NSAIDs)

Propensity score results using a 1:1 matched sample are provided in the Appendix C, Table 4c. Analyses run with either lookwise score estimates or cumulative score estimates failed to signal at any of the four time points. In contrast, the disease risk scores matched 1:1 signaled slightly earlier than the 1:4 matched analyses (Appendix C, Table 5c). One reason for the difference in the disease risk score pattern could be related to differences in the proximity of matches for 1:1 versus 1:4 matches. Table 19 shows estimates of the differences in propensity score or disease risk scores for the 1:1 and 1:4 matched

cohorts. When the calipers were initially set, examinations of the disease risk score showed a range of near 0 up to 0.7 similar to the propensity score range of near 0 to just over 0.8. Subsequently, disease risk scores at one site were found to be problematic and the final disease risk score results removed those scores. The remaining disease risk scores were lower (maximum 0.43 and large proportion of scores < 0.01) and a smaller caliper to restrict matches further might have been useful.

Table 19. Differences in scores for matched samples (table shows estimates from lookwise estimation)

Score difference between case and control(s)	PS[1] score: 1:1 match (N=80532 COX-2) (N=80532 N-NSAID)	PS score 4:1 match (N=80532 COX-2) (N=234456 N-NSAID)	DRS[2] score 1:1 match (N=79524 COX-2) (N=79524 N-NSAID)	DRS score 4:1 match (N=79524 COX-2) (N=317979 N-NSAID)
0	71140 (88.34%)	258308 (82.01%)	72272 (90.88%)	267048 (83.98%)
<0.001	9078 (11.27%)	50401 (16.00%)	7227 (9.09%)	50168 (15.78%)
0.001 to < 0.01	287 (0.36%)	3130 (0.99%)	22 (0.03%)	596 (0.19%)
0.01 to < 0.02	17 (0.02%)	1016 (0.32%)	3 (<0.01%)	83 (0.03%)
0.02 to < 0.03	8 (0.01%)	819 (0.26%)	0	45 (0.01%)
0.03 to < 0.04	2 (<0.01%)	700 (0.22%)	0	26 (0.01%)
0.04 to < 0.05	0	614 (0.19%)	0	13 (<0.01%)

[1]PS=propensity score; [2]Disease Risk Score (scores based on results from two sites due to DRS scores problems at 1 site)

Stratification Adjusted Analyses

We explored deciles of propensity scores or disease risk scores as potential adjustments using stratified analysis. Stratified adjustment using deciles of propensity scores for the full cohort are shown in Table 20. Both lookwise and cumulative estimates of the propensity scores fail to reach a signal. Disease risk scores results similarly did not produce a significant positive test statistic (Table 21). Differing distributions of the propensity scores by COX-2 versus non-selective NSAID users are evident and are one issue in these naïve analyses (Table 22). It is typical practice to restrict propensity score analyses to subjects whose scores are in a range where more equipoise is evident. However, in this example, restricting by deciles would remove a large number of the rare outcomes (Table 23) and subsequent analyses lose power. It is less intuitive to truncate disease risk scores and such a step would obviously restrict outcome numbers as well. While it is natural to look within strata to evaluate differential risks there will be less power within strata.

Table 20. Incident GI bleed outcome: results for Propensity Score stratified analyses (10 strata)

Look	Time[1]	Nonselective NSAID		COX-2 NSAID		Adjusted HR[4] Nonselective NSAID vs COX-2	Sequential analysis test statistic[5]	Test statistic boundary[6]	Signal[7]
		GI bleeds$_{cum}$[2]	Person years$_{cum}$[3]	GI bleeds$_{cum}$	Person years$_{cum}$				
Lookwise PS[8]									
1	2008	866	275934	63	11510	1.07	0.573	2.1040	No
2	2009	1942	628523	130	22824	1.05	0.534	2.0323	No
3	2010	2787	920050	174	30703	1.05	0.633	2.0514	No
4	2011	3544	1186492	224	36878	0.98	-0.253	2.0508	No
Cumulative PS[9]									
1	2008	866	275934	63	11510	0.96	-0.277	2.1040	No
2	2009	1942	628523	130	22824	0.96	-0.499	2.0323	No
3	2010	2787	920050	174	30703	0.96	-0.540	2.0514	No
4	2011	3544	1186492	224	36878	0.89	-1.625	2.0508	No

Table 21. Incident GI bleed outcome: results for Disease Risk Score stratified analyses (10 strata)

Look	Time[1]	Nonselective NSAID		COX-2 NSAID		HR[4] Nonselective NSAID vs COX-2	Sequential analysis test statistic[5]	Test statistic boundary[6]	Signal[7]
		GI bleeds$_{cum}$[2]	Person years$_{cum}$[3]	GI bleeds$_{cum}$	Person years$_{cum}$				
Lookwise DRS[8]									
1	2008	813	225170	63	11403	1.05	0.403	2.1040	No
2	2009	1842	531931	130	22602	1.04	0.397	2.0323	No
3	2010	2647	781430	173	30366	1.04	0.526	2.0514	No
4	2011	3365	1005950	220	36388	0.99	-0.206.	2.0508	No

Look	Time[1]	Nonselective NSAID		COX-2 NSAID		HR[4] Nonselective NSAID vs COX-2	Sequential analysis test statistic[5]	Test statistic boundary[6]	Signal[7]
		GI bleeds$_{cum}$[2]	Person years$_{cum}$[3]	GI bleeds$_{cum}$	Person years$_{cum}$				
Cumulative DRS[9]									
1	2008	813	225170	63	11403	1.06	0.420	2.1040	No
2	2009	1842	531931	130	22602	1.04	0.392	2.0323	No
3	2010	2647	781430	173	30366	1.04	0.500	2.0514	No
4	2011	3365	1005950	220	36388	0.98	-0.237	2.0508	No

[1]Timing of each look; Later looks include information from the preceding time intervals

[2]GI bleeds$_{cum}$ is the total number of incident GI bleed events up to and including the year of each look (within each selected drug group) , only two sites included due to DRS data issues at third site

[3]Person years$_{cum}$ is the total accumulated exposure time up to and including the year of each look (within each selected drug group), only two sites included due to DRS data issues at third site

[4]HR is the hazard ratio for nonselective NSAID vs COX-2 drugs for all times up to and including each look

[5]Test Statistic is the drug comparison parameter estimate / standard error from the proportional hazards model

[6]Test statistic boundary: sequential analysis boundary estimate

[7]Signal indicates when the test statistic first exceeds the test statistic boundary

[8]Lookwise estimation used single year data when estimating scores

[9]Cumulative estimation used all years up to and including the current year when estimating scores

Table 22. COX-2 and non-selection NSAID use by propensity score strata

PROPENSITY SCORE DECILES (LOOKWISE ESTIMATION)	NON-SELECTIVE NSAID	COX-2
1	241361 (9.3%)	28102 (34.9%)
2	256151 (9.8%)	14044 (17.4%)
3	262307 (10.1%)	9132 (11.3%)
4	261565 (10.0%)	6317 7.8%
5	270638 (10.4%)	5842 (7.3%)
6	265054 (10.2%)	4793 (6.0%)
7	263431 (10.1%)	3999 (5.0%)
8	269597 (10.3%)	3191 (4.0%)
9	263032 (10.1%)	2714 (3.4%)
10	255231 (9.8%)	2401 (3.0%)

Table 23. GI Bleed outcome by propensity score strata

	GI BLEED WITHIN 6 MONTHS	
PROPENSITY SCORE DECILES (LOOKWISE ESTIMATION)	**No**	**Yes**
1	268235 (10.0%)	1228 (32.6%)
2	269623 (10.0%)	572 (15.2%)
3	271018 (10.1%)	421 (11.2%)
4	267470 (10.0%)	412 10.9%
5	276223 (10.3%)	257 (6.8%)
6	269640 (10.0%)	207 (5.5%)
7	267261 (10.0%)	169 (4.5%)
8	272651 (10.2%)	137 (3.6%)
9	265600 (9.9%)	146 (3.9%)
10	257414 (9.6%)	218 (5.8%)

IV. DISCUSSIONS

A. DISCUSSION OF SIMULATION RESULTS

Our simulation results showed that empirical power increased with higher matching ratios, whether based on propensity scores or based on disease risk scores, especially when both simulated treatments and simulated outcomes were rare. While a high matching ratio (e.g., 1:8) preserved more cases in the analytic dataset and increased empirical power, in real data applications, a higher matching ratio may not be feasible due to a limited pool of appropriate controls. For matching, we matched to the nearest score without a caliper because it would not be feasible to choose an optimal caliper for each simulated dataset in a simulation setting with different combinations of parameters. However, preliminary examination showed that the matching resulted in matched pairs with small differences between scores, particularly in rare settings.

In this simulation work, we used PS or DRS scores in stratifying current year incident users into 10 strata. In general, stratification performed less well compared to 1:4 when treatment and outcome were rare while all methods were comparable when treatment and outcome were common. This stratification approach kept all cases in the analytic datasets, but it may not be the optimal method when adjusting for confounders in sequential analyses in observational studies. Specifically, we observed that stratification based on either PS or DRS scores inflated type I error rates and empirical power, especially

when both outcomes and treatments were rare. Using this stratification approach, very different persons may be put in a stratum. Finer stratification (more strata) potentially could improve the performance of stratification as more strata more closely simulate the matching approach. Since the stratification method included all cases in the analyses, more datasets that yielded valid treatment estimates would be expected. However, that was not what we observed in early years of monitoring (Table 8). In addition, due to the limitations such as using asymptotic Lan-Demets Group Sequential approach in calculating boundaries and not considering delaying of the first analysis, further exploration of stratification is required.

In matched settings, when we had fewer, delayed looks (i.e. biennial rather than annual), there was slightly higher empirical power and type I error rates. However, changes were not large and were not evident in the stratified analyses. In analyzing real observational data, if treatments and outcomes are evenly distributed over the follow-up time, as in the simulation setting, researchers may conduct sequential analyses as scheduled given the minimal power gains with a delay. However, in analyzing real data, if there is not more information at a scheduled look, researchers may consider skipping that look until a later time when more data are available.

Instead of using individuals in the current year in the lookwise approach, our cumulative approach used data from all individuals from the beginning of monitoring to the current year to estimate PS or DRS scores. Because it includes more individuals, the cumulative approach can accommodate more confounders. In addition, cumulative methods may produce scores with less uncertainty (smaller standard error for predicted scores) due to a larger sample size. However, because the process of matching and stratification use only predicted PS/DRS scores and do not typically account for the uncertainty of these scores, the cumulative approach did not differ from the lookwise approach.[15] The other reason for benefit of applying the cumulative approach not apparent in the simulation setting is that those performance measurements were averaged across 1000 replicates. Although the cumulative approach did not demonstrate an advantage here, we recommend using it in safety monitoring of emergent treatment because the lookwise approach may require exclusion of important confounders (due to non-convergence in building PS or DRS models) because of few outcomes or treatments.

We encountered non-convergence when building PS and DRS models when treatments and outcomes were rare (always) or moderately rare (sometimes). To solve this problem in a simulation setting, we chose backward covariate selection method to have the SAS codes run without manually selecting covariates for each dataset. The selection of covariates with $p<0.5$ to be included in the final model is the level Frank Harrell recommends.[16] The primary issues relate to problems of p values being too liberal and models being 'data driven' so that they won't replicate in other settings etc. However, the goal of this simulation study is to compare different approaches in adjusting for confounders on average based on 1000 replicates NOT based on each individual dataset. Thus, we believe the use of backward selection of covariates in building PS or DRS scores is appropriate in the simulation setting.

We applied the Lan-Demets Group Sequential approach in this simulation study. This approach is an approximate, not an exact approach (i.e., the Monte Carlo approach).[17] The stopping boundaries based on this approach used an asymptotic theory, which may not hold for the scenario of rare treatment and rare outcome. It may be one of the causes to inflated type I error rates in regression and stratification approaches. However, we applied the same boundaries to all approaches, yielding relative and consistent comparisons. The test statistics based on the Lan-Demets Group Sequential approach were

shown in Table A5. The trends of these test statistics supported the observations of type I error rates and empirical powers in the simulation study. Boundaries were not modified when counts of individuals and outcomes differed after score matching or stratification. We purposely calculated the upper boundaries using eligible subjects and number of cases, and then applied the same boundaries to different methods.

No changes in the ability to adjust for confounding between DRS and PS methods were observed when confounding strengths changed. We did not observe either an advantage of the DRS approach when the association of confounders with outcomes was stronger than the association with treatment or an advantage of the PS approach when the association of confounders with treatment was stronger than the association with outcomes. The DRS approach may be less useful when there are few outcomes. Similarly, PS approach may be less useful when there are few treatments.[4] PS has been advocated as advantageous when a single PS can be used for multiple outcomes, however disparate outcomes may require different exclusions or variable adjustments that limit duplicate use. Similarly, a single DRS model may be advantageous when studying multiple exposures but only if the population for the DRS model is appropriate for the differing exposures. Although available, methods for creating PS are not trivial when there are multiple exposures or with multiple exposure levels. In such instances, the DRS method may be advantageous because a single DRS model can be fit by including categorical variables for the exposure variable.

Invalid estimates of treatment effect were more likely due to the fact that there was no outcomes in the treatment group than in the comparator group because we set up the simulation such that the majority of the population was comparators. In calculating type I error rates, datasets with invalid estimates were considered not signaled. When treatments and outcomes were rare, the average of valid treatment effect estimates was overestimated, which was consistent with previous studies. [18,19,20] When treatments and outcomes were common, more datasets produced valid estimates and the average of valid estimates was unbiased.

The same information including number of subjects, number of outcomes and confounders was used for DRS and PS approaches. We did not observe significant difference between PS and DRS approaches in type I error rates and empirical powers. A limitation of this simulation study was that we did not simulate the situation where historical data could be used in building DRS model. Another limitation of this simulation study is that we did not simulate a multi-site study as in the NSAID example.

It is possible for the scenario of rare treatment and rare outcome that there is no reasonable amount of information. Another limitation is that we did not include delaying of the first analysis because of practical reasons. The SAS programs for conducting the simulation study are very complex, each containing over 9000 SAS statements. Each SAS program can generate confounders, simulate treatment status, simulate outcome, build PS and DRS models (stepwise and cumulative), perform 1:1 and 1:4 matching and stratification, and conduct sequential analyses. Inclusion of delaying the first analyses will further complicate the SAS programs and lengthen the running time of these programs which already took dozens of hours to run.

B. DISCUSSION OF THE METHODOLOGICAL EXAMPLE

The goal of comparing the risk of GI bleeding after incident use of COX-2 and nonselective NSAID drugs was to examine similar comparative methods in a data application in order to gain additional insights.

Similar to the simulation study, we did not observe meaningful differences in results when scores were created using lookwise or cumulative estimation methods. Operationally, the lookwise method has an advantage in not requiring historical data when estimating scores at later looks. Although one might expect the cumulative method to have more stable scores, neither the simulation nor the example runs displayed evidence of this potential advantage.

In comparisons of matching methods and stratification, the NSAID example agreed with the simulation results in suggesting matching as the preferred method. The example further highlighted how difficult it is to adequately adjust for confounding using stratification in a setting where scores have disparate distributions for the drug comparisons of interest.

Unlike the simulation, the NSAID example showed some differences in results for comparisons of propensity scores and disease risk scores as the disease risk score had slightly earlier or stronger signals. However, this single example does not support generalizations since the simulations would have had some runs where each score prevailed despite the overall similarities "on average". Nevertheless, the varying results on the example are helpful in highlighting several issues that will likely occur in data applications. Unlike in the simulation setting where both the propensity score models and the disease risk score models are equally correct, in real world settings we often don't know the "truth" and one score may unknowingly lack critical covariates. We can also expect the distributions of the two different scores rarely to be similar and balancing across the groups of interest will likely differ. In most applied settings, there are advantages to having both scores calculated and reviewed. If both methods signal, it will add some level of reassurance to the results. If only one method signals, additional data checks and other efforts to understand the differences will ensure a deeper understanding of any results prior to dissemination.

C. LIMITATIONS

One advantage of the disease risk score is that historical data might be available to build the model. The simulation study did not directly examine such a scenario but it is likely that results where treatments were rare but outcomes were moderate or common may reflect results that might be expected in such an instance. Another limitation is that the simulation study did not mimic the more complicated multi-site study setting that was used in the NSAID example. Nor did the simulation study use a survival outcome, but instead assumed a binary outcome without censoring. Therefore some issues of using survival outcomes should be further explored. The NSAID example was limited in not employing a variety of covariate inclusion methods (i.e. a small consistent model and a full model to be evaluated with the selected models).

V. CONCLUSION AND RECOMMENDATIONS

1. This study found that stratification methods performed less well than matching methods, which is consistent with the finding reported by Austin et al (2007). [22] Matched cohort methods are preferable to stratification to adjust for potential confounders.
2. For rare outcomes and rare treatments, we recommend that, whenever feasible, a high matching ratio (e.g., 1:8) be used to maximize statistical power as long as matched score differences don't exceed the selected caliper estimate. For common and moderately rare

outcomes and treatments, a lower matching ratio (e.g., 1:4) is sufficient in maximizing statistical power in signaling detection.

3. Use both propensity score and disease risk score adjustments if possible.
4. Although the cumulative approach did not demonstrate an advantage here, we recommend using it in safety monitoring of emergent treatment because the lookwise approach may require exclusion of important confounders (due to non-convergence in building PS or DRS models) because of few outcomes or treatments.

VI. REFERENCES

1. Li L, Kulldorff M, Nelson JC, Cook AJ. A Propensity Score-Enhanced Sequential Analytic Method for Comparative Drug Safety Surveillance. Statistics in Biosciences 2011; 3: 45-62.
2. Miettinen OS. Stratification by a multivariate confounder score. Am J Epidemiol. 1976; 104:609–620.
3. Arbogast PG, Ray WA. Performance of disease risk scores, propensity scores, and traditional multivariable outcome regression in the presence of multiple confounders. Am J Epidemiol. 2011; 174:613–620.
4. Glynn RJ, Gagne JJ, and Schneeweiss S. Role of disease risk scores in comparative effectiveness research with emerging therapies. Pharmacoepidemiol Drug Saf. 2012; 21(Suppl 2): 138–147.
5. Liang KY. Extended Mantel-Haenszel estimating procedure for multivariate logistic regression models. Biometrics 1987; 43(2):289-99.
6. Firth D. Bias reduction of maximum likelihood estimates. Biometrika 1993; 80, 27.
7. Molenberghs G, Verbeke G: Models for Discrete Longitudinal Data. Berlin, Springer; 2005.
8. Lan KKG and DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika 1983, 70:659-63.
9. Brookhart AM, Wang PS, Solomon DH and Schneeweiss S. Evaluating short term drug effects using a physician-specific prescribing preference as an instrumental variable. Epidemiology 2006; 17: 268-275.
10. 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; 20:512-522.
11. Go As, Hylek EM, Chang Y et al. Anitcoagulation therapy for stroke prevention in atrial fibrillation: how well do randomized trials translate into clinical practice? JAMA 2003; 290:2685-2692. (Note: ICD codes not in published article, provided by study authors.)
12. Cattaruzzi C, Troncon MG, Agostinis L. et al. Positive predictive value ofICD-9th codes for upper gastrointestinal bleeding and perforation in the Sistema Informativo Sanitario Regionale Database. J. Clinical Epidemiol. 1999; 52(6): 499-502.
13. Raiford DS, Gutthann SP, Rodriguez LAG. 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.
14. Quan H, Sundararajan V, Halfon P et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Medical Care 2005; 43: 1130-1139.
15. McCandless LC, Gustafson P, Austin PC. Bayesian propensity score analysis for observational data. Stat Med. 2009; 28(1):94-112. doi: 10.1002/sim.3460.
16. Harrell F E. Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis, Springer-Verlag, New York. 2001.
17. Shao J and Feng H. Group sequential t-test for clinical trials with small sample sizes across stages. Contemporary Clinical Trials 2007; 28: 563–571.
18. King G and Zeng L. Logistic Regression in Rare Events Data. Political Analysis 2001; 9: 137–163.

19. Glanz JM, McClure DL, Xu S, Hambidge SJ, Lee M, Kolczak MS. et al. Four different study designs to evaluate vaccine safety were equally validated with contrasting limitations. Journal of clinical epidemiology 2006; 59: 808-818.
20. Zeng C, Newcomer SR, Glanz JM, Shoup JA, Daley MF, Hambidge SJ, and Xu S. Bias Correction of Risk Estimates in Vaccine Safety Studies With Rare Adverse Events Using a Self-controlled Case Series Design. American Journal of Epidemiology 2013. doi: 10.1093/aje/kwt211
21. Pocock SJ. Interim Analyses for randomized clinical-trials - The Group Sequential Approach. Biometrics. 1982; 38: 153-62.
22. Austin PC, Grootendorst P and Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Statist. Med. 2007; 26:734–753

VII. APPENDIX A

Table A1. Intercept parameter coefficients used in simulating datasets (N=5000 in each year)

	OUTCOME INTERCEPT				TREATMENT INTERCEPT			
	y	Scenario 1	Scenario 2	Scenario 3	n*	Scenario 1	Scenario 2	Scenario 3
Rare	<20/5000	-6.0	-6.0	-6.85	~25/5000	-6.08	-6.91	-6.08
Moderate	~50/5000	-4.8	-4.8	-5.63	~50/5000	-5.25	-6.25	-5.40
Common	~225-/5000	-3.3	-3.3	-4.15	>500/5000	-2.5	-3.3	-2.5

*Treatment n is for 1st look, new treatment becomes more common over time

In tables A2-A6, scenarios 1-3 represent the following
Scenario 1: Strengths of associations between confounders and exposure and confounders and outcome are the same
Scenario 2: Strength of association between confounders and exposure is greater
Scenario 3: Strength of association between confounders and outcome is greater

Table A2. Comparison of different approaches for analyzing stratified data based on propensity scores when the treatment of interest is commonly prescribed: number of datasets with valid treatment effects and mean treatment effect (standard deviation) for treatment effects θ_Z= 0.69

SCENARIO*	θ_Z	$\bar{n}$	$\bar{y}$	STRATA	NUMBER OF DATASETS WITH VALID TREATMENT EFFECT ESTIMATES AND MEAN TREATMENT EFFECT (STD)				
					Mantel-Haenszel	Firth	Conditional	Fixed effects	Random effect
1	0.69	459	17	5	414 -7.16 (2.95)	1000 1.08 (0.40)	946 0.65 (0.62)	946 0.65 (0.62)	946 0.67 (0.62)
				10	96 -9.22 (2.11)	1000 1.30 (0.31)	946 0.65 (0.62)	946 0.65 (0.62)	946 0.67 (0.62)
				20	11 -10.55 (1.70)	1000 1.43 (0.26)	946 0.65 (0.62)	946 0.65 (0.62)	946 0.67 (0.62)
2	0.69	463	17	5	408 -7.47 (3.03)	1000 1.28 (0.38)	958 0.67 (0.64)	958 0.67 (0.64)	958 0.72 (0.64)
				10	111 -9.95 (1.80)	1000 1.39 (0.31)	958 0.67 (0.64)	958 0.67 (0.64)	958 0.72 (0.63)
				20	10 -11.07 (0.72)	1000 1.50 (0.27)	958 0.66 (0.64)	958 0.67 (0.65)	958 0.72 (0.63)
3	0.69	459	17	5	458 -7.47 (3.16)	1000 1.11 (0.40)	949 0.66 (0.63)	949 0.66 (0.63)	949 0.71 (0.63)
				10	127 -9.25 (2.34)	1000 1.28 (0.32)	949 0.66 (0.63)	949 0.66 (0.63)	949 0.71 (0.63)
				20	18 -9.83 (3.29)	1000 1.40 (0.27)	949 0.66 (0.63)	949 0.66 (0.63)	958 0.71 (0.63)
1	0.69	463	109	5	999 -0.26 (2.01)	1000 0.68 (0.28)	1000 0.67 (0.27)	1000 0.67 (0.27)	1000 0.68 (0.27)
				10	996 -3.51 (2.86)	1000 0.73 (0.25)	1000 0.66 (0.27)	1000 0.67 (0.27)	1000 0.68 (0.27)
				20	853 -8.21 (2.11)	1000 0.90 (0.20)	1000 0.66 (0.27)	1000 0.67 (0.27)	1000 0.68 (0.27)
2	0.69	463	109	5	1000 -1.22 (2.71)	1000 0.70 (0.29)	1000 0.68 (0.27)	1000 0.68 (0.27)	1000 0.73 (0.27)
				10	993 -4.70 (2.71)	1000 0.79 (0.25)	1000 0.67 (0.27)	1000 0.67 (0.27)	1000 0.73 (0.27)
				20	792 -8.74 (1.93)	1000 0.99 (0.20)	1000 0.67 (0.28)	1000 0.67 (0.28)	1000 0.73 (0.27)
3	0.69	459	108	5	1000 -0.97 (2.64)	1000 0.70 (0.29)	1000 0.68 (0.27)	1000 0.68 (0.27)	1000 0.70 (0.27)
				10	990 -4.49 (2.98)	999 0.78 (0.25)	999 0.68 (0.27)	999 0.68 (0.27)	999 0.70 (0.27)
				20	825 -8.664 (2.02)	999 0.95 (0.20)	999 0.67 (0.27)	999 0.68 (0.27)	999 0.71 (0.27)

$\bar{n}$, mean number of the treatment of interest; $\bar{y}$, mean number of outcomes.

Table A3. Comparing different approaches for analyzing stratified data based on propensity scores when treatment of interest is rarely prescribed: number of datasets with valid treatment effects and mean treatment effect (standard deviation) for treatment effect θ_Z= 0.69

SCENARIO*	θ_Z	$\bar{n}$	$\bar{y}$	STRATA	NUMBER OF DATASETS WITH VALID TREATMENT EFFECT ESTIMATES AND MEAN TREATMENT EFFECT (STD)				
					Mantel-Haenszel	Firth	Conditional	Fixed effects	Random effect
1	0.69	25	16	5	46 -9.07 (1.19)	124 3.42 (0.56)	148 2.67 (0.42)	148 2.67 (0.43)	148 2.67 (0.40)
				10	10 -9.67 (2.19)	37 3.79 (2.11)	148 2.68 (0.45)	148 2.70 (0.46)	148 2.67 (0.40)
				20	41 -8.00 (3.80)	57 3.29 (0.42)	148 2.69 (0.47)	148 2.73 (0.49)	148 2.66 (0.40)
2	0.69	26	16	5	32 -8.54 (2.74)	100 3.26 (0.93)	103 2.63 (0.43)	103 2.64 (0.44)	103 2.66 (0.41)
				10	9 -8.54 (3.80)	13 3.69 (0.28)	104 2.54 (1.25)	103 2.68 (0.53)	103 2.67 (0.41)
				20	23 -8.49 (2.07)	41 3.31 (0.38)	103 2.63 (0.47)	103 2.66 (0.49)	103 2.65 (0.42)
3	0.69	25	16	5	40 -8.85 (1.32)	126 3.35 (0.70)	149 2.64 (0.44)	149 2.65 (0.45)	149 2.63 (0.44)
				10	34 -7.85 (2.40)	88 3.50 (0.48)	149 2.64 (0.45)	149 2.66 (0.46)	149 2.63 (0.44)
				20	40 -8.01 (2.23)	56 3.25 (0.42)	150 2.59 (1.07)	149 2.70 (0.52)	149 2.63 (0.44)
1	0.69	25	100	5	312 -8.61 (3.57)	544 1.82 (0.50)	659 1.04 (0.46)	659 1.04 (0.46)	659 1.04 (0.46)
				10	44 -10.14 (1.46)	79 3.25 (2.22)	659 1.04 (0.46)	659 1.04 (0.47)	659 1.04 (0.46)
				20	132 -8.25 (3.77)	269 1.84 (0.87)	659 1.04 (0.47)	659 1.03 (0.47)	659 1.04 (0.46)
2	0.69	26	101	5	222 -9.27 (2.40)	458 2.05 (0.50)	708 1.00 (0.50)	708 1.01 (0.50)	708 1.06 (0.49)
				10	60 -10.14(1.52)	58 3.03 (1.93)	708 1.00 (0.50)	708 1.00 (0.50)	708 1.05 (0.49)
				20	17 -11.03 (0.89)	0	708 1.00 (0.51)	708 1.00 (0.51)	708 1.05 (0.49)
3	0.69	25	100	5	319 -8.03 (4.11)	532 1.83 (0.58)	649 1.08 (0.49)	649 1.08 (0.49)	649 1.07 (0.49)
				10	188 -6.99 (4.29)	365 2.17 (1.55)	649 1.08 (0.49)	649 1.08 (0.50)	649 1.06 (0.49)
				20	147 -7.70 (3.83)	254 1.81 (0.82)	649 1.07 (0.495)	649 1.08 (0.50)	649 1.06 (0.49)

$\bar{n}$, mean number of the treatment of interest; $\bar{y}$, mean number of outcomes.

Table A4. Mean (standard deviation) of treatment effect by look with $\theta_z = 0.69$

SCENARIOS	LOOK	PROPENSITY SCORE			DISEASE RISK SCORE		
		Matching 1:1	Matching 1:4	Stratification	Matching 1:1	Matching 1:4	Stratification
1	1	4.73 (11.91)	0.02 (4.12)	-0.051 (3.23)	4.52 (11.80)	0.02 (4.14)	-0.052 (3.24)
	2	1.63 (5.60)	0.56 (1.50)	0.54 (1.32)	1.48 (5.28)	0.56 (1.50)	0.54 (1.33)
	3	0.93 (2.35)	0.67 (0.41)	0.66 (0.38)	1.07 (2.85)	0.66 (0.42)	0.65 (0.39)
	4	0.74 (0.50)	0.68 (0.32)	0.68 (0.30)	0.78 (0.52)	0.68 (0.33)	0.67 (0.30)
	5	0.73 (0.43)	0.69 (0.27)	0.68 (0.29)	0.75 (0.43)	0.68 (0.28)	0.68 (0.26)
	6	0.72 (0.37)	0.68 (0.25)	0.67 (0.24)	0.73 (0.38)	0.67 (0.25)	0.67 (0.24)
	7	0.72 (0.34)	0.68 (0.23)	0.67 (0.21)	0.73 (0.33)	0.68 (0.23)	0.68 (0.21)
	8	0.72 (0.31)	0.68 (0.21)	0.68 (0.20)	0.73 (0.30)	0.68 (0.21)	0.68 (0.20)
	9	0.71 (0.29)	0.69 (0.19)	0.68 (0.18)	0.72 (0.28)	0.69 (0.19)	0.68 (0.18)
	10	0.71 (0.27)	0.69 (0.17)	0.68 (0.17)	0.72 (0.26)	0.68 (0.18)	0.68 (0.17)
2	1	4.13 (10.95)	0.14 (3.17)	0.14 (3.17)	4.05 (10.88)	0.20 (3.42)	0.19 (2.79)
	2	1.32 (4.30)	0.64 (0.98)	0.64 (0.98)	1.79 (5.43)	0.65 (0.96)	0.65 (0.87)
	3	0.84 (1.71)	0.68 (0.40)	0.68 (0.40)	1.05 (2.84)	0.69 (0.39)	0.70 (0.37)
	4	0.73 (0.49)	0.69 (0.32)	0.68 (0.32)	0.87 (1.69)	0.70 (0.32)	0.71 (0.30)
	5	0.72 (0.41)	0.69 (0.29)	0.69 (0.29)	0.75 (0.45)	0.71 (0.28)	0.70 (0.27)
	6	0.70 (0.37)	0.69 (0.26)	0.69 (0.25)	0.73 (0.39)	0.71 (0.26)	0.70 (0.25)
	7	0.71 (0.34)	0.70 (0.23)	0.70 (0.23)	0.72 (0.33)	0.71 (0.23)	0.71 (0.22)
	8	0.70 (0.31)	0.71 (0.22)	0.71 (0.21)	0.72(0.30)	0.72 (0.21)	0.71 (0.21)
	9	0.70 (0.28)	0.71 (0.20)	0.72(0.20)	0.72 (0.28)	0.73 (0.19)	0.71 (0.19)
	10	0.70 (0.25)	0.72 (0.18)	0.72 (0.18)	0.72 (0.25)	0.73 (0.18)	0.71 (0.18)
3	1	4.78 (11.70)	0.12 (4.14)	0.02 (3.13)	4.77 (11.61)	0.07 (3.79)	0.05 (3.13)
	2	1.71 (5.45)	0.60 (1.21)	0.58 (1.07)	1.68 (5.35)	0.62 (1.21)	0.61 (1.07)
	3	0.84 (1.63)	0.66 (0.42)	0.65 (0.39)	0.93 (2.23)	0.69 (0.43)	0.68 (0.39)
	4	0.73 (0.49)	0.68 (0.33)	0.67 (0.30)	0.85 (1.60)	0.70 (0.34)	0.70 (0.30)
	5	0.73 (0.43)	0.68 (0.29)	0.68 (0.27)	0.75 (0.48)	0.71 (0.30)	0.71 (0.27)
	6	0.71 (0.37)	0.68 (0.26)	0.67 (0.24)	0.74 (0.41)	0.70 (0.27)	0.70 (0.24)
	7	0.72 (0.34)	0.68 (0.23)	0.68 (0.22)	0.74 (0.38)	0.71 (0.24)	0.70 (0.22)
	8	0.72 (0.30)	0.69 (0.21)	0.69 (0.20)	0.74 (0.33)	0.71 (0.21)	0.71 (0.20)
	9	0.71 (0.28)	0.70 (0.19)	0.69 (0.18)	0.73 (0.31)	0.72 (0.20)	0.72 (0.19)
	10	0.70 (0.26)	0.71 (0.18)	0.69 (0.18)	0.73 (0.28)	0.72 (0.19)	0.72 (0.18)

Table A5. Mean of test statistic by look with θ_z=0.69

SCENARIOS	LOOK	PROPENSITY SCORE			DISEASE RISK SCORE		
		Matching 1:1	Matching 1:4	Stratification	Matching 1:1	Matching 1:4	Stratification
1	1	0.37	0.94	1.06	0.35	0.94	1.06
	2	0.89	1.44	1.58	0.89	1.44	1.59
	3	1.18	1.78	1.93	1.21	1.76	1.94
	4	1.47	2.13	2.29	1.51	2.11	2.30
	5	1.70	2.44	2.60	1.72	2.42	2.61
	6	1.90	2.69	2.87	1.92	2.67	2.88
	7	2.12	2.98	3.15	2.13	2.96	3.17
	8	2.34	3.28	3.44	2.46	3.27	3.46
	9	2.53	3.57	3.71	2.55	3.56	3.73
	10	2.72	3.84	3.96	2.74	3.83	3.98
2	1	0.42	1.04	1.13	0.42	1.03	1.19
	2	0.92	1.52	1.63	0.94	1.53	1.73
	3	1.24	1.86	1.97	1.24	1.87	2.08
	4	1.50	2.20	2.30	1.53	2.23	2.44
	5	1.72	2.52	2.61	1.76	2.55	2.75
	6	1.91	2.81	2.87	1.96	2.85	3.04
	7	2.15	3.14	3.17	2.17	3.17	3.35
	8	2.34	3.47	3.45	2.39	3.50	3.64
	9	2.54	3.78	3.71	2.60	3.81	3.92
	10	2.75	4.07	3.94	2.80	4.10	4.17
3	1	0.37	0.96	1.09	0.36	1.00	1.14
	2	0.88	1.44	1.58	0.90	1.48	1.65
	3	1.20	1.75	1.91	1.19	1.80	2.00
	4	1.44	2.09	2.26	1.46	2.15	2.36
	5	1.68	2.41	2.59	1.70	2.49	2.71
	6	1.88	2.67	2.85	1.91	2.75	2.98
	7	2.11	2.98	3.14	2.13	3.06	3.29
	8	2.31	3.28	3.41	2.35	3.37	3.57
	9	2.49	3.60	3.69	2.56	3.69	3.85
	10	2.69	3.90	3.94	2.76	3.99	4.12

Table A6a. Rare treatment and rare outcome, lookwise estimation method, scenario 1 after increasing treatment parameter to 1.61*

TREATMENT PARAMETER θ_Z=1.61 (OR=5)					
Design and Method	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
		Signaled	**Signaled and Not, N**		
PS Matching 1:1	8.1	2.18 (0.27)	1.35 (0.65), 746	9.59 (0.61)	10 (8 10)
PS Matching 1:4	78.2	2.10 (0.59)	1.83 (0.77), 995	6.96 (1.90)	7 (4 10)
PS Stratification	90.3	2.10 (0.63)	1.96 (0.74), 998	4.10 (2.70)	3 (1 9)
DRS Matching 1:1	8.1	2.22 (0.26)	1.35 (0.68), 743	9.48 (0.71)	10 (8 10)
DRS Matching 1:4	78.4	2.10 (0.60)	1.82 (0.78), 996	6.96 (1.94)	7 (4 10)
DRS Stratification	90.4	2.10 (0.63)	1.97 (0.74), 998	4.10 (2.70)	3 (1 9)

*Description of the columns in Tables6a-6c, see the second paragraph of Appendix B.

Table A6b. Rare treatment rare outcome, lookwise estimation method, scenario 2 after increasing treatment parameter to 1.61

TREATMENT PARAMETER θ_Z=1.61 (OR=5)					
Design and Method	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
		Signaled	**Signaled and Not, N**		
PS Matching 1:1	0.5	1.34 (0.05)	1.03 (0.67), 504	10 (NA)	10 (10, 10)
PS Matching 1:4	46.5	2.24 (0.53)	1.71 (0.80), 893	7.96 (1.68)	8 (5 10)
PS Stratification	77.3	2.33 (0.70)	2.07 (0.88), 924	4.51 (2.95)	4 (1 10)
DRS Matching 1:1	0.5	2.08 (0.35)	1.05 (0.65), 534	10 (NA)	10 (10, 10)
DRS Matching 1:4	46.6	2.27 (0.53)	1.72 (0.81), 889	8.00 (1.75)	8 (5 10)
DRS Stratification	76.8	2.32 (0.69)	2.06 (0.88), 924	4.48 (2.97)	4 (1 10)

Table A6c. Rare treatment rare outcome, lookwise estimation method, scenario 3 after increasing treatment parameter to 1.61

TREATMENT PARAMETER θ_Z=1.61 (OR=5)					
Design and Method	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
		Signaled	**Signaled and Not, N**		
PS Matching 1:1	0.3	2.13 (0.29)	0.91 (0.71), 517	9.67 (0.58)	10 (9 10)
PS Matching 1:4	45.9	2.29 (0.51)	1.68 (0.84), 931	8.02 (1.67)	8 (5 10)
PS Stratification	77.9	2.39 (0.73)	2.06 (0.95), 974	4.44 (3.01)	4 (1 10)
DRS Matching 1:1	0.2	2.30 (NA)	0.93 (0.70), 513	9.00 (NA)	9 (9 9)
DRS Matching 1:4	44.9	2.34 (0.52)	1.70 (0.85), 927	8.11 (1.63)	8 (5 10)
DRS Stratification	78.3	2.39 (0.73)	2.07 (0.95), 974	4.43 (3.01)	4 (1 10)

VIII. APPENDIX B

This appendix provides results from all simulations that were performed on an annual basis for a period of ten years. The treatment parameter is set to 0 and 0.69. In the tables below, "lookwise estimation method" and "cumulative estimation method" refer to the method for creating PS model and DRS model in the sequential analyses (Section II, 6a and 6b). "Scenario 1", "scenario 2", and "scenario 3" refer to different scenarios in which the strengths of associations between confounders $x_1 - x_4$ and exposure and confounders $x_1 - x_4$ and outcome differ (Table 1).

Description of the columns:
Type 1 error %: Percentage of false positives when the treatment parameter is set to 0.
Signaled %: Empirical power when the treatment parameter is set to 0.69.
Signaled: Average treatment parameter obtained from signaled replicates when the treatment parameter is set to 0.69.
Signaled and not: Average parameter obtained from signaled replicates and those that did not signal at the end of follow-up (only datasets that converged were included) when the treatment parameter is set to 0.69.
N = Number of replicate data sets that converged i.e., the number that was used to calculate the average parameter

We included all variables listed in Table 1 to compute PSs and DRSs used in these analyses with the exception of scenarios in which either the treatment or outcome was rare or moderately rare. In order to ensure that the models converged, we used a backward selection method (p<0.5) in these scenarios to limit the number of variables included.

Table B1-1. Rare treatment and rare outcome, lookwise estimation method, and scenario 1

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1)**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
	Type 1 Error %		**Signaled**	**Signaled and Not, N**		
PS Matching 1:1	0.0	0.0	NA	0.40 (0.71), 712	NA	NA
PS Matching 1:4	2.0	14.7	1.94 (0.48)	0.74 (0.80), 929	8.19 (1.68)	8 (5 10)
PS Stratification	10.4	36.4	1.91 (0.66)	0.93 (0.96), 932	4.95 (3.15)	5 (1 10)
DRS Matching 1:1	0.0	0.0	NA	0.44 (0.75), 711	NA	NA
DRS Matching 1:4	1.6	13.3	1.94 (0.45)	0.73 (0.78), 930	8.11 (1.70)	8 (5 10)
DRS Stratification	10.1	36.0	1.93 (0.66)	0.93 (0.96), 929	4.84 (3.13)	4 (1 10)

Table B1-2. Rare treatment and rare outcome, lookwise estimation method, and scenario

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0	0.0	NA	0.23 (0.63), 711	NA	NA
PS Matching 1:4	0.9	5.1	2.19 (0.48)	0.85 (0.77), 799	8.67 (1.70)	9 (5 10)
PS Stratification	11	30.6	2.24 (0.76)	1.14 (1.01), 818	4.86 (3.29)	5 (1 10)
DRS Matching 1:1	0	0.0	NA	0.22 (0.62), 569	NA	NA
DRS Matching 1:4	0.8	4.7	2.18 (0.44)	0.84 (0.79), 805	8.70 (1.63)	9 (6 10)
DRS Stratification	10.7	30.6	2.25 (0.75)	1.14 (1.02), 818	4.82 (3.32)	5 (1 10)

Table B1-3. Rare treatment and rare outcome, lookwise estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0	0.0	NA	0.19 (0.60), 528	NA	NA
PS Matching 1:4	1.0	5.8	2.30 (0.45)	0.92 (0.81), 762	8.52 (1.50)	9 (6 10)
PS Stratification	10.9	30.8	2.38 (0.77)	1.26 (1.06), 784	4.45 (3.19)	4 (1 10)
DRS Matching 1:1	0	0.0	NA	0.22 (0.57), 526	NA	NA
DRS Matching 1:4	0.5	5.7	2.29 (0.45)	0.92 (0.80), 755	8.42 (1.69)	9 (5 10)
DRS Stratification	11.6	31.0	2.37 (0.75)	1.27 (1.05), 783	4.40 (3.17)	4 (1 10)

Table B1-4. Rare treatment and rare outcome, cumulative estimation method, and scenario 1

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0	0.1	2.30 (NA)	0.40 (0.75), 714	9.0 (NA)	9.0 (NA)
PS Matching 1:4	1.7	14.9	2.01 (0.50)	0.75 (0.82), 932	8.05 (1.70)	8 (5 10)
PS Stratification	9.9	35.8	1.93 (0.66)	0.93 (0.96), 932	4.84 (3.09)	4 (1 10)
DRS Matching 1:1	0	0.1	2.30 (NA)	0.36 (0.74), 724	9.0 (NA)	9.0 (NA)
DRS Matching 1:4	1.7	14.8	1.92 (0.46)	0.73 (0.80), 928	8.34 (1.63)	8.5 (5 10)
DRS Stratification	10.7	36.0	1.93 (0.66)	0.94 (0.96), 932	4.78 (3.10)	4 (1 10)
regression	10	35.9	1.93 (0.67)	1.50 (0.78), 572	4.81 (3.08)	4 (1 10)

Table B1-5. Rare treatment and rare outcome, cumulative estimation method, and scenario 2

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0	0	NA	0.23 (0.62), 536	NA	NA
PS Matching 1:4	0.6	5.5	2.31 (0.49)	0.84 (0.76), 802	8.84 (1.71)	9 (4 10)
PS Stratification	11.1	30.3	2.25 (0.78)	1.13 (1.01), 818	4.87 (3.30)	5 (1 10)
DRS Matching 1:1	0	0	NA	0.22 (0.63), 553	NA	NA
DRS Matching 1:4	0.7	4.9	2.13 (0.46)	0.86 (0.78), 792	8.61 (1.51)	9 (6 10)
DRS Stratification	10.9	31.0	2.24 (0.75)	1.15 (1.01), 818	4.87 (3.34)	5 (1 10)
regression	10.2	28.7	2.25 (0.77)	1.66 (0.89), 507	4.80 (3.24)	5 (1 10)

Table B1-6. Rare treatment and rare outcome, cumulative estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0	0	NA	0.26 (0.58), 496	NA	NA
PS Matching 1:4	0.5	6.6	2.24 (0.47)	0.94 (0.79), 758	8.41 (1.68)	9 (5 10)
PS Stratification	11.3	31.0	2.38 (0.77)	1.27 (1.05), 785	4.41 (3.17)	4 (1 10)
DRS Matching 1:1	0	0	NA	0.22 (0.60), 515	NA	NA
DRS Matching 1:4	0.7	6.3	2.23 (0.51)	0.92 (0.81), 759	8.67 (1.51)	9 (6 10)
DRS Stratification	11.5	30.9	2.37 (0.75)	1.24 (1.13), 786	4.45 (3.20)	4 (1 10)
regression	10.9	30.1	2.35 (0.78)	1.84 (0.92), 474	4.56 (3.20)	4 (1 10)

Table B2-1. Rare treatment and moderate outcome, lookwise estimation method, and scenario 1

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0.0	1.0	2.26 (0.32)	0.68 (0.73), 921	9.50 (0.53)	9.50 (9, 10)
PS Matching 1:4	3.8	28.2	1.75 (0.55)	0.79 (0.79), 993	7.19 (2.05)	7 (3, 10)
PS Stratification	8.7	44.8	1.57 (0.54)	0.88 (0.79), 993	5.18 (2.95)	5 (1, 10)
DRS Matching 1:1	0.0	1.3	2.06 (0.37)	0.72 (0.74), 912	9.69 (0.48)	10 (9, 10)
DRS Matching 1:4	4.6	30.9	1.72 (0.50)	0.82 (0.77), 992	7.18 (1.97)	7 (3, 10)
DRS Stratification	9.0	45.2	1.58 (0.55)	0.89 (0.79), 993	5.13 (2.94)	5 (1 10)

Table B2-2. Rare treatment and moderate outcome, lookwise estimation method, and scenario2

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0	2.1	2.11 (0.31)	0.74 (0.72), 929	9.33 (0.80)	10 (8, 10)
PS Matching 1:4	4.3	28.8	1.68 (0.50)	0.81 (0.72), 995	7.03 (2.11)	7 (3, 10)
PS Stratification	8.7	47.4	1.51 (0.52)	0.91 (0.73), 995	5.40 (2.97)	5 (1, 10)
DRS Matching 1:1	0	2.0	2.15 (0.33)	0.71 (0.71), 917	9.60 (0.60)	10 (8.5, 10)
DRS Matching 1:4	4.1	30.4	1.69 (0.49)	0.82 (0.73), 995	7.23 (2.10)	7 (4, 10)
DRS Stratification	9.3	49.4	1.51 (0.52)	0.93 (0.74), 995	5.44 (2.98)	6 (1, 10)

Table B2-3. Rare treatment and moderate outcome, lookwise estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0	1.4	2.19 (0.29)	0.68 (0.75), 929	9.71 (0.61)	10 (8, 10)
PS Matching 1:4	3.9	30.3	1.67(0.51)	0.79 (0.77), 997	7.16 (2.12)	7 (3, 10)
PS Stratification	10.5	48.5	1.53 (0.53)	0.89 (0.79), 997	5.28 (2.85)	6 (1, 10)
DRS Matching 1:1	0	1.5	2.13 (0.29)	0.69 (0.78), 915	9.67 (0.49)	10 (9, 10)
DRS Matching 1:4	3.8	30.8	1.69 (0.52)	0.80 (0.78), 997	7.21 (2.07)	7 (3, 10)
DRS Stratification	10.1	49.1	1.53 (0.52)	0.90 (0.79), 997	5.27 (2.84)	5 (1, 10)

Table B2-4. Rare treatment and moderate outcome, cumulative estimation method, scenario 1

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0.0	1.6	2.16 (0.28)	0.72 (0.77), 911	9.38 (0.62)	9.00 (8, 10)
PS Matching 1:4	3.8	28.9	1.75 (0.53)	0.80 (0.79), 993	7.16 (2.02)	7 (3, 10)
PS Stratification	8.7	44.8	1.57 (0.55)	0.88 (0.79), 993	5.21 (2.93)	5 (1, 10)
DRS Matching 1:1	0.0	1.7	2.15 (0.29)	0.69 (0.75), 907	9.41 (0.62)	9 (8, 10)
DRS Matching 1:4	4.1	29.9	1.72 (0.53)	0.81 (0.77), 993	7.37 (1.99)	8 (4, 10)
DRS Stratification	9.4	46.0	1.58 (0.55)	0.90 (0.79), 993	5.18 (2.91)	5 (1, 10)
Regression	8.7	44.9	1.57 (0.55)	0.87 (0.89), 994	5.16 (2.96)	5 (1, 10)

Table B2-5. Rare treatment and moderate outcome, cumulative estimation method, and scenario2

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0	1.6	2.16 (0.32)	0.73 (0.71), 923	9.31 (0.87)	10 (8, 10)
PS Matching 1:4	4.4	28.3	1.68 (0.50)	0.80 (0.75), 995	7.17 (2.02)	8 (3, 10)
PS Stratification	8.3	45.4	1.51 (0.53)	0.88 (0.74), 995	5.37 (2.98)	5 (1, 10)
DRS Matching 1:1	0	2.3	2.15 (0.29)	0.77 (0.72), 910	9.43 (0.73)	10 (8, 10)
DRS Matching 1:4	3.9	29.0	1.74 (0.53)	0.85 (0.75), 995	7.10 (2.05)	7 (3, 10)
DRS Stratification	9.6	49.4	1.52 (0.52)	0.94 (0.74), 995	5.38 (2.95)	6 (1, 10)
Regression	7.8	45.7	1.51 (0.53)	0.87 (0.84), 996	5.36 (3.00)	5 (1, 10)

Table B2-6. Rare treatment and moderate outcome, cumulative estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0	1.6	2.08 (0.31)	0.67 (0.76), 923	9.75 (0.58)	10 (8, 10)
PS Matching 1:4	4.5	30.0	1.70 (0.54)	0.78 (0.79), 997	7.16 (2.06)	7 (3, 10)
PS Stratification	10.2	47.0	1.54 (0.54)	0.87 (0.80), 997	5.28 (2.89)	5 (1, 10)
DRS Matching 1:1	0.1	1.6	1.99 (0.33)	0.69 (0.75), 938	9.75 (0.45)	10 (9, 10)
DRS Matching 1:4	4.7	30.3	1.74 (0.54)	0.80 (0.80), 997	7.10 (2.12)	7 (3, 10)
DRS Stratification	9.9	47.4	1.54 (0.53)	0.89 (0.79), 997	5.24 (2.85)	5 (1, 10)
Regression	10.1	46.4	1.54 (0.56)	0.83 (1.06), 1000	5.23 (2.91)	5 (1, 10)

Table B3-1. Rare treatment and common outcome, lookwise estimation method, and scenario 1

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	1.6	42.0	1.29 (0.44)	0.81 (0.53),1000	7.70 (1.67)	8 (5 10)
PS Matching 1:4	5.6	72.6	1.03 (0.41)	0.85 (0.47),1000	6.18 (2.52)	6 (2 10)
PS Stratification	7.9	82.7	0.99 (0.39)	0.87 (0.43),1000	5.24 (2.76)	5 (1 10)
DRS Matching 1:1	1.6	48.2	1.30 (0.45)	0.86 (0.56),1000	7.70 (1.63)	8 (5 10)
DRS Matching 1:4	5.0	73.3	1.08 (0.44)	0.89 (0.50), 1000	5.95 (2.48)	6 (2 10)
DRS Stratification	8.4	83.7	0.99 (0.38)	0.88 (0.43), 1000	5.23 (2.74)	5 (1, 10)

Table B3-2. Rare treatment and common outcome, lookwise estimation method, and scenario 2

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	2.2	45.2	1.31 (0.44)	0.84 (0.55),1000	7.68 (1.66)	8 (5 10)
PS Matching 1:4	5.7	74.0	1.11 (0.47)	0.92 (0.52),1000	5.86 (2.53)	6 (2 10)
PS Stratification	9.4	84.9	1.03 (0.39)	0.93 (0.44),999	4.91 (2.70)	5 (1 10)
DRS Matching 1:1	1.6	47.7	1.34 (0.46)	0.89 (0.56),1000	7.58 (1.70)	8 (5 10)
DRS Matching 1:4	2.2	75.4	1.12 (0.48)	0.94 (0.53),1000	5.84 (2.54)	6 (2 10)
DRS Stratification	3.8	86.3	1.04 (0.39)	0.95 (0.44),1000	4.86 (2.65)	5 (1 9)

Table B3-3: Rare treatment and common outcome, lookwise estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	2.2	45.8	1.30 (0.42)	0.84 (0.54),1000	7.47 (1.68)	7.5 (5 10)
PS Matching 1:4	5.9	74.9	1.08 (0.44)	0.91 (0.50),1000	5.87 (2.54)	6 (2 10)
PS Stratification	8.2	86.2	1.00 (0.39)	0.91 (0.43),1000	5.11 (2.72)	5 (1 10)
DRS Matching 1:1	3.2	48.1	1.29 (0.42)	0.86 (0.53),1000	7.50 (1.70)	7 (5 10)
DRS Matching 1:4	5.4	76.4	1.08 (0.44)	0.91 (0.49),1000	5.85 (2.52)	6 (2 10)
DRS Stratification	8.6	86.1	1.00 (0.40)	0.91 (0.44),1000	5.13 (2.71)	5 (1 10)

Table B3-4. Rare treatment and common outcome, cumulative estimation method, and scenario 1

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	1.7	44.4	1.29 (0.43)	0.83 (0.53),1000	7.72 (1.67)	8 (5 10)
PS Matching 1:4	5.9	72.7	1.07 (0.44)	0.89 (0.39),1000	5.97 (2.48)	6 (2 10)
PS Stratification	8.1	83.5	0.99 (0.38)	0.88 (0.43),1000	5.25 (2.76)	5 (1 10)
DRS Matching 1:1	3.0	46.7	1.34 (0.44)	0.87 (0.56),1000	7.59 (1.67)	8 (5 10)
DRS Matching 1:4	6.2	73.0	1.07 (0.44)	0.88 (0.50),1000	6.03 (2.51)	6 (2 10)
DRS Stratification	8.8	83.6	0.99 (0.39)	0.88 (0.43),1000	5.23 (2.76)	5 (1 10)
regression	7.8	82.8	0.98 (0.39)	0.87 (0.44),1000	5.29 (2.77)	5 (1 10)

Table B3-5. Rare treatment and common outcome, cumulative estimation method, and scenario 2

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	2.3	45.5	1.33 (0.43)	0.87 (0.55), 1000	7.57 (1.71)	8 (5 10)
PS Matching 1:4	6.5	75.3	1.10 (0.46)	0.92 (0.51), 1000	5.89 (2.53)	6 (2 10)
PS Stratification	9.5	84.6	1.03 (0.39)	0.92 (0.44), 1000	4.92 (2.70)	5 (1 10)
DRS Matching 1:1	2.4	47.3	1.32 (0.44)	0.88 (0.54), 1000	7.64 (1.75)	8 (5 10)
DRS Matching 1:4	5.4	75.6	1.13 (0.48)	0.95 (0.53), 1000	5.78 (2.56)	6 (2 10)
DRS Stratification	10.4	85.7	1.04 (0.39)	0.94 (0.44), 1000	4.87 (2.68)	5 (1 9)
regression	8.4	82.9	1.00 (0.39)	0.89 (0.44), 1000	5.10 (2.73)	5 (1 10)

Table B3-6. Rare treatment and common outcome, cumulative estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	1.8	48.2	1.29 (0.44)	0.86 (0.54), 1000	7.57 (1.74)	8 (5 10)
PS Matching 1:4	7.1	77.5	1.07 (0.43)	0.92 (0.49), 1000	5.91 (2.52)	6 (2 10)
PS Stratification	9.4	87.2	1.00 (0.39)	0.92 (0.43), 1000	5.09 (2.69)	5 (1 10)
DRS Matching 1:1	3.2	48.2	1.32 (0.43)	0.88 (0.55), 1000	7.44 (1.68)	7 (5 10)
DRS Matching 1:4	5.4	76.4	1.08 (0.44)	0.91 (0.49), 1000	5.91 (2.55)	6 (2 10)
DRS Stratification	8.2	85.9	1.00 (0.39)	0.91 (0.44), 1000	5.13 (2.72)	5 (1 10)
regression	7.2	84.5	1.00 (0.40)	0.89 (0.45), 1000	5.18 (2.74)	5 (1 10)

Table B4-1. Moderate treatment and rare outcome, lookwise estimation method, and scenario 1

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0	0.7	2.16 (0..32)	0.60 (0.75), 872	8.86 (0.38)	10 (9 10)
PS Matching 1:4	2.8	24.8	1.76 (0.51)	0.74 (0.79), 989	7.62 (2.08)	8 (4 10)
PS Stratification	9.0	39.7	1.62 (0.55)	0.82 (0.82), 989	5.26 (2.96)	5 (1 10)
DRS Matching 1:1	0	0.6	2.23 (0.26)	0.58 (0.77), 881	9.83 (0.41)	10 (9 10)
DRS Matching 1:4	3.1	24.3	1.82 (0.54)	0.72 (0.83), 988	7.52 (2.04)	8 (4 10)
DRS Stratification	9.2	38.9	a.62 (0.54)	0.81 (0.82), 988	5.22 (2.92)	5 (1 10)

Table B4-2. Moderate treatment-rare outcome, lookwise estimation method, and scenario 2

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0	1.5	2.08 (0.32)	0.69 (0.76), 915	9.33 (1.05)	10 (7 10)
PS Matching 1:4	3.4	30.6	1.73 (0.55)	0.81 (0.79), 995	7.23 (2.02)	7 (4 10)
PS Stratification	8.6	44.4	1.54 (0.53)	0.88 (0.76), 995	5.19 (2.90)	5 (1 10)
DRS Matching 1:1	0	1.5	2.06 (0.39)	0.70 (0.76), 928	9.20 (0.86)	9 (8 10)
DRS Matching 1:4	3.5	29.7	1.74 (0.52)	0.81 (0.78), 995	7.08 (2.02)	7 (4 10)
DRS Stratification	8.6	44.2	1.54 (0.52)	0.87 (0.76), 995	5.21 (2.90)	5 (1 10)

Table B4-3. Moderate treatment and rare outcome, lookwise estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**			**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**	**Mean Time To Signal Detection (STD)**	
PS Matching 1:1	0	0.8	2.10 (0.40)	0.67 (0.75), 912	9.63 (0.74)	10 (8 10)
PS Matching 1:4	3.7	28.1	1.76 (0.54)	0.81 (0.78), 992	7.22 (2.08)	8 (4 10)
PS Stratification	9.0	46.6	1.54 (0.52)	0.90 (0.77), 992	5.23 (2.93)	5 (1 10)
DRS Matching 1:1	0	0.7	2.17 (0.32)	0.65 (0.75), 904	9.57 (0.79)	10 (8 10)
DRS Matching 1:4	3.8	27.3	1.76 (0.55)	0.79 (0.78), 992	7.15 (2.13)	7 (3 10)
DRS Stratification	8.1	45.1	1.56 (0.53)	0.89 (0.77), 992	5.26 (2.93)	5 (1 10)

Table B4-4. Moderate treatment and rare outcome, cumulative estimation method, and scenario 1

Design and Method	θ_Z=0 (OR=1) Type 1 Error %	Signaled %	MEAN $\hat{\theta}_Z$ (STD) Signaled	MEAN $\hat{\theta}_Z$ (STD) Signaled and Not, N	Mean Time to Signal Detection (STD)	Median Time to Signal Detection (5%, 95%)
PS Matching 1:1	0	0.6	2.23 (0.22)	0.57 (0.75), 869	9.50 (0.84)	10 (8 10)
PS Matching 1:4	3.1	23.8	1.79 (0.53)	0.72 (0.81), 989	7.46 (1.97)	8 (4 10)
PS Stratification	9.0	39.5	1.62 (0.55)	0.83 (0.82), 989	5.24 (2.95)	5 (1 10)
DRS Matching 1:1	0	0.4	2.32 (0.05)	0.58 (0.78), 880	9.50 (1.00)	10 (8 10)
DRS Matching 1:4	3.6	24.2	1.80 (0.52)	0.73 (0.80), 989	7.52 (1.98)	8 (4 10)
DRS Stratification	8.7	39.1	1.62 (0.54)	0.82 (0.82), 989	5.22 (2.91)	5 (1 10)
regression	8.3	39.2	1.61 (0.56)	1.06 (0.80), 729	5.32 (2.96)	5 (1 10)

Table B4-5. Moderate treatment and rare outcome, cumulative estimation method, and scenario 2

	TREATMENT PARAMETER θ_Z=0.69 (OR=2)					
Design and Method	θ_Z=0 (OR=1) Type 1 Error %	Signaled %	MEAN $\hat{\theta}_Z$ (STD) Signaled	MEAN $\hat{\theta}_Z$ (STD) Signaled and Not, N	Mean Time to Signal Detection (STD)	Median Time to Signal Detection (5%, 95%)
PS Matching 1:1	0	1.4	2.23 (0.26)	0.69 (0.76), 922	9.36 (1.08)	10 (7 10)
PS Matching 1:4	2.6	29.2	1.68 (0.51)	0.78 (0.76), 995	7.27 (1.98)	8 (4 10)
PS Stratification	7.5	44.4	1.53 (0.53)	0.87 (0.76), 995	5.24 (2.92)	5 (1 10)
DRS Matching 1:1	0	1.7	2.14 (0.30)	0.70 (0.74), 916	9.41 (0.94)	10 (7 10)
DRS Matching 1:4	2.9	2.90	1.76 (0.53)	0.81 (0.78), 995	7.03 (2.08)	7 (4 10)
DRS Stratification	7.5	45.0	1.53 (0.53)	0.88 (0.76), 995	5.31 (2.92)	5 (1 10)
regression	6.8	43.7	1.51 (0.52)	1.05 (0.71), 784	5.35 (2.94)	5 (1 10)

Table B4-6. Moderate treatment and rare outcome, cumulative estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0	0.8	2.20 (0.31)	0.66 (0.75), 906	9.12 (0.83)	9 (8 10)
PS Matching 1:4	4.3	28.5	1.72 (0.50)	0.80 (0.76), 992	7.28 (2.05)	8 (4 10)
PS Stratification	8.6	46.1	1.55 (0.52)	0.90 (0.77), 992	5.28 (2.92)	5 (1 10)
DRS Matching 1:1	0	1.2	2.19 (0.36)	0.67 (0.76), 904	9.67 (0.65)	10 (8 10)
DRS Matching 1:4	3.7	27.6	1.74 (0.53)	0.78 (0.78), 992	7.29 (2.11)	8 (3 10)
DRS Stratification	8.4	45.2	1.55 (0.53)	0.88 (0.77), 992	5.24 (2.91)	5 (1 10)
regression	7.7	43.0	1.55 (0.54)	1.07 (0.73), 777	5.31 (2.95)	5 (1 10)

Table B5-1. Moderate treatment and moderate outcome, lookwise estimation method, and scenario 1

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0.7	19.7	1.77 (0.43)	0.83 (0.64), 997	8.70 (1.14)	9 (7 10)
PS Matching 1:4	5.8	50.9	1.39 (0.55)	0.90 (0.66), 1000	6.31 (2.41)	6 (2 10)
PS Stratification	8.6	65.7	1.24 (0.45)	0.92 (0.59), 1000	5.21 (2.96)	5 (1 10)
DRS Matching 1:1	0.5	17.2	1.76 (0.43)	0.79 (0.64), 992	8.67 (1.22)	9 (7 10)
DRS Matching 1:4	5.2	49.6	1.37 (0.54)	0.88 (0.66), 1000	6.40 (2.45)	7 (3 10)
DRS Stratification	8.8	65.9	1.24 (0.46)	0.93 (0.59), 1000	5.24 (2.97)	5 (1 10)

Table B5-2. Moderate treatment and moderate outcome, lookwise estimation method, and scenario 2

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0.7	16.9	1.79 (0.39)	0.83 (0.65), 993	8.82 (1.07)	9 (7 10)
PS Matching 1:4	5.9	49.4	1.43 (0.54)	0.91 (0.67),1000	6.33 (2.43)	6 (3 10)
PS Stratification	9.7	66.8	1.28 (0.48)	0.97 (0.61),1000	5.16 (3.02)	5 (1 10)
DRS Matching 1:1	0.8	16.1	1.83 (0.44)	0.83 (0.65), 995	8.71 (1.22)	9 (7 10)
DRS Matching 1:4	6.8	50.9	1.38 (0.52)	0.91 (0.64),1000	6.54 (2.45)	7 (2 10)
DRS Stratification	9.9	67.5	1.28 (0.48)	0.98 (0.61),1000	5.17 (2.99)	5 (1 10)

Table B5-3. Moderate treatment and moderate outcome, lookwise estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0.2	15.8	1.79 (0.44)	0.80 (0.66), 995	8.86 (1.11)	9 (7 10)
PS Matching 1:4	4.5	50.0	1.40 (0.51)	0.89 (0.66),1000	6.46 (2.43)	7 (2.5 10)
PS Stratification	8.5	65.4	1.28 (0.46)	0.96 (0.60),1000	5.06 (2.86)	5 (1 10)
DRS Matching 1:1	0.2	16.1	1.84 (0.44)	0.80 (0.67), 998	8.83 (1.12)	9 (7 10)
DRS Matching 1:4	4.5	50.9	1.37 (0.50)	0.89 (0.65),1000	6.55 (2.38)	7 (2 10)
DRS Stratification	8.4	65.2	1.27 (0.46)	0.95 (0.60),1000	5.13 (2.86)	5 (1 10)

Table B5-4. Moderate treatment and moderate outcome, cumulative estimation method, and scenario 1

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0.6	18.3	1.73 (0.45)	0.79 (0.63),999	8.72 (1.20)	9 (7 10)
PS Matching 1:4	5.0	50.3	1.37 (0.51)	0.89 (0.64),1000	6.32 (2.47)	6 (2 10)
PS Stratification	8.7	65.7	1.24 (0.45)	0.92 (0.59),1000	5.20 (2.95)	5 (1 10)
DRS Matching 1:1	0.6	17.7	1.80 (0.46)	0.80 (0.65),996	8.74 (1.21)	9 (7 10)
DRS Matching 1:4	5.7	49.1	1.39 (0.54)	0.88 (0.66),1000	6.32 (2.47)	7 (2 10)
DRS Stratification	8.8	65.6	1.24 (0.46)	0.93 (0.59),999	5.20 (2.96)	5 (1 10)
regression	8.4	65.1	1.24 (0.46)	1.03 (0.54),883	5.20 (2.96)	5 (1 10)

Table B5-5: Moderate treatment and moderate outcome, cumulative estimation method, and scenario 2

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	0.6	15.5	1.83 (0.42)	0.81 (0.68),994	8.72 (1.14)	9 (7 10)
PS Matching 1:4	4.7	50.4	1.39 (0.52)	0.90 (0.65),1000	6.46 (2.49)	7 (2 10)
PS Stratification	9.8	66.2	1.29 (0.48)	0.96 (0.61),1000	5.12 (3.02)	5 (1 10)
DRS Matching 1:1	0.6	16.2	1.85 (0.45)	0.83 (0.66),987	8.65 (1.17)	9 (7 10)
DRS Matching 1:4	6.4	50.5	1.42 (0.57)	0.92 (0.68),1000	6.46 (2.44)	7 (2 10)
DRS Stratification	10.3	67.4	1.28 (0.48)	0.98 (0.61),1000	5.13 (2.97)	5 (1 10)
regression	8.9	64.2	1.27 (0.48)	1.04 (0.56),895	5.17 (3.00)	5 (1 10)

Table B5-6. Moderate treatment and moderate outcome, cumulative estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1)**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
	Type 1 Error %		**Signaled**	**Signaled and Not, N**		
PS Matching 1:1	0.5	16.1	1.79 (0.45)	0.81 (0.65), 997	8.86 (1.13)	9 (7 10)
PS Matching 1:4	4.9	50.3	1.40 (0.52)	0.89 (0.67),1000	6.43 (2.43)	6 (3 10)
PS Stratification	8.9	65.6	1.28 (0.46)	0.96 (0.60),1000	5.08 (2.85)	5 (1 10)
DRS Matching 1:1	0.3	17.0	1.83 (0.43)	0.79 (0.68),994	8.88 (1.08)	9 (7 10)
DRS Matching 1:4	5.2	50.5	1.41 (0.53)	0.90 (0.67),1000	6.47 (2.47)	7 (2 10)
DRS Stratification	8.7	63.9	1.28 (0.46)	0.94 (0.61),1000	5.05 (2.82)	5 (1 10)
regression	7.8	61.9	1.27 (0.46)	1.02 (0.55),883	5.15 (2.90)	5 (1 10)

Table B6-1. Moderate treatment and common outcome, lookwise estimation method, and scenario 1

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1)**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
	Type 1 Error %		**Signaled**	**Signaled and Not, N**		
PS Matching 1:1	4.2	77.4	1.06 (0.45)	0.91 (0.49), 1000	6.49 (2.20)	6 (3 10)
PS Matching 1:4	6.2	93.4	0.93 (0.41)	0.89 (0.43), 1000	4.86 (2.38)	5 (1 9)
PS Stratification	8.2	97.6	0.87 (0.32)	0.86 (0.33), 1000	4.14 (2.34)	4 (1 9)
DRS Matching 1:1	2.8	78.0	1.06 (0.47)	0.91 (0.50), 1000	6.54 (2.14)	7 (3 10)
DRS Matching 1:4	6.8	94.0	0.95 (0.41)	0.91 (0.43), 1000	4.77 (2.40)	4 (1 9)
DRS Stratification	6.8	97.0	0.88 (0.32)	0.86 (0.34), 1000	4.11 (2.30)	4 (1 9)

Table B6-2. Moderate treatment and common outcome, lookwise estimation method, and scenario 2

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	4.8	89.0	0.98 (0.46)	0.92 (0.48), 1000	5.94 (2.30)	6 (3 10)
PS Matching 1:4	8.8	97.6	0.93 (0.43)	0.92 (0.43), 1000	4.11 (2.23)	4 (1 9)
PS Stratification	9.4	99.4	0.86 (0.30)	0.86 (0.31), 1000	3.44 (1.99)	3 (1 7)
DRS Matching 1:1	6.2	91.6	0.99 (0.46)	0.94 (0.47), 1000	5.86 (2.19)	6 (3 10)
DRS Matching 1:4	10.0	98.0	0.92 (0.39)	0.91 (0.40), 1000	4.11 (2.16)	4 (1 8)
DRS Stratification	10.8	99.4	0.89 (0.30)	0.89 (0.31), 1000	3.30 (1.92)	3 (1 7)

Table B6-3. Moderate treatment and common outcome, lookwise estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	4.4	83.2	1.02 (0.43)	0.92 (0.46), 1000	6.18 (2.14)	6 (3 10)
PS Matching 1:4	7.4	98.0	0.94 (0.47)	0.93 (0.47), 1000	4.68 (2.40)	4 (1 9)
PS Stratification	11.4	99.2	0.89 (0.32)	0.89 (0.32), 1000	3.75 (2.14)	3 (1 8)
DRS Matching 1:1	4.2	82.6	1.05 (0.46)	0.93 (0.49), 1000	6.15 (2.12)	6 (3 10)
DRS Matching 1:4	8.0	97.0	0.94 (0.40)	0.92 (0.40), 1000	4.48 (2.21)	4 (1 8)
DRS Stratification	10.6	99.0	0.89 (0.32)	0.88 (0.33), 1000	3.79 (2.19)	3 (1 7)

Table B6-4. Moderate treatment and common outcome, cumulative estimation method, and scenario 1

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	2.8	76.6	1.06 (0.44)	0.91 (0.48), 1000	6.47 (2.15)	6 (3 10)
PS Matching 1:4	6.4	94.0	0.94 (0.40)	0.90 (0.42), 1000	4.86 (2.37)	5 (1 9)
PS Stratification	8.0	96.8	0.87 (0.33)	0.85 (0.34), 1000	4.10 (2.29)	4 (1 9)
DRS Matching 1:1	4.0	76.8	1.07 (0.44)	0.91 (0.49), 1000	6.38 (2.06)	6 (3 10)
DRS Matching 1:4	7.2	92.4	0.96 (0.42)	0.91 (0.44), 1000	4.78 (2.46)	4 (1 9)
DRS Stratification	7.8	97.0	0.87 (0.32)	0.86 (0.34), 1000	4.10 (2.28)	4 (1 8)
regression	8.0	96.4	0.87 (0.33)	0.85 (0.34), 1000	4.13 (2.31)	4 (1 9)

Table B6-5. Moderate treatment and common outcome, cumulative estimation method, and scenario 2

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	4.4	90.8	0.99 (0.44)	0.93 (0.46), 1000	5.87 (2.21)	6 (3 10)
PS Matching 1:4	8.8	97.4	0.94 (0.43)	0.93 (0.43), 1000	4.10 (2.25)	4 (1 8)
PS Stratification	8.8	99.4	0.86 (0.30)	0.86 (0.31), 1000	3.47 (2.02)	3 (1 7)
DRS Matching 1:1	5.8	91.8	0.97 (0.41)	0.92 (0.43), 1000	5.88 (2.16)	6 (3 10)
DRS Matching 1:4	8.0	97.8	0.93 (0.39)	0.91 (0.40), 1000	4.08 (2.18)	4 (1 9)
DRS Stratification	10.0	99.4	0.88 (0.31)	0.88 (0.31), 1000	3.34 (1.96)	3 (1 7)
regression	7.8	99.2	0.84 (0.30)	0.84 (0.31), 1000	3.60 (2.10)	3 (1 8)

Table B6-6. Moderate treatment and common outcome, cumulative estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	4.0	83.2	1.02 (0.43)	0.91 (0.46), 1000	6.26 (2.17)	6 (3 10)
PS Matching 1:4	7.6	97.6	0.94 (0.44)	0.92 (0.45), 1000	4.61 (2.35)	4 (1 9)
PS Stratification	11.0	99.2	0.90 (0.32)	0.89 (0.32), 1000	3.72 (2.11)	3 (1 8)
DRS Matching 1:1	2.8	82.6	1.01 (0.42)	0.90 (0.45), 998	6.20 (2.09)	6 (3 10)
DRS Matching 1:4	7.2	97.6	0.93 (0.39)	0.92 (0.39), 998	4.56 (2.33)	4 (1 9)
DRS Stratification	10.6	98.8	0.88 (0.32)	0.88 (0.33), 998	3.80 (2.19)	4 (1 8)
regression	8.6	98.8	0.87 (0.33)	0.86 (0.33), 998	3.95 (2.25)	4 (1 8)

Table B7-1. Common treatment and rare outcome, lookwise estimation method, and scenario 1

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	3.7	78.6	1.04 (0.42)	0.89 (0.47), 1000	6.24 (2.15)	6 (3 10)
PS Matching 1:4	6.7	95.5	0.95 (0.44)	0.92 (0.45), 1000	4.56 (2.43)	4 (1 9)
PS Stratification	7.7	96.2	0.91 (0.37)	1.05 (0.46), 1000	4.22 (2.49)	4 (1 9)
DRS Matching 1:1	3.0	79.1	1.05 (0.46)	0.91 (0.50), 996	6.27 (2.21)	6 (3 10)
DRS Matching 1:4	6.6	95.7	0.93 (0.42)	0.91 (0.43), 996	4.64 (2.44)	4 (1 9)
DRS Stratification	7.9	95.8	0.91 (0.36)	0.89 (0.37), 995	4.19 (2.48)	4 (1 9)

Table B7-2. Common treatment and rare outcome, lookwise estimation method, and scenario 2

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	2.5	81.1	1.00 (0.42)	0.88 (0.46), 1000	6.29 (2.20)	6 (3 10)
PS Matching 1:4	7.1	96.7	0.92 (0.40)	0.90 (0.41), 1000	4.56 (2.42)	4 (1 9)
PS Stratification	6.6	95.7	0.91 (0.37)	0.88 (0.38), 1000	4.26 (2.47)	4 (1 9)
DRS Matching 1:1	2.9	80.1	1.04 (0.45)	0.90 (0.49), 1000	6.21 (2.21)	6 (3 10)
DRS Matching 1:4	8.7	96.7	0.94 (0.41)	0.92 (0.42), 1000	4.44 (2.36)	4 (1 9)
DRS Stratification	7.5	96.6	0.92 (0.36)	0.90 (0.38), 1000	4.05 (2.40)	4 (1 9)

Table B7-3. Common treatment and rare outcome, lookwise estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	2.5	76.4	1.04 (0.42)	0.88 (0.47), 1000	6.22 (2.17)	6 (3 10)
PS Matching 1:4	6.2	95.5	0.94 (0.41)	0.91 (0.42), 1000	4.65 (2.46)	4 (1 9)
PS Stratification	6.1	95.9	0.92 (0.36)	0.89 (0.37), 1000	4.21 (2.45)	4 (1 9)
DRS Matching 1:1	3.5	78.8	1.06 (0.44)	0.90 (0.48), 1000	6.23 (2.15)	6 (3 10)
DRS Matching 1:4	5.9	96.3	0.98 (0.46)	0.95 (0.47), 1000	4.48 (2.46)	4 (1 9)
DRS Stratification	7.4	96.6	0.94 (0.37)	0.92 (0.38), 1000	4.04 (2.39)	4 (1 9)

Table B7-4. Common treatment and rare outcome, lookwise estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	2.9	79.2	1.03 (0.41)	0.89 (0.45), 1000	6.32 (2.15)	6 (3 10)
PS Matching 1:4	6.3	95.4	0.94 (0.44)	0.91 (0.45), 1000	4.58 (2.45)	4 (1 9)
PS Stratification	7.9	96.3	0.91 (0.37)	0.89 (0.38), 1000	4.22 (2.49)	4 (1 9)
DRS Matching 1:1	3.2	78.7	1.05 (0.46)	0.91 (0.49), 1000	6.31 (2.17)	6 (3 10)
DRS Matching 1:4	6.7	96.2	0.93 (0.42)	0.90 (0.43), 1000	4.62 (2.43)	4 (1 9)
DRS Stratification	8.1	96.7	0.91 (0.36)	0.89 (0.37), 1000	4.21 (2.49)	4 (1 9)
Regression	7.1	96.5	0.91 (0.36)	0.89 (0.37), 999	4.23 (2.47)	4 (1 9)

Table B7-5. Common treatment and rare outcome, cumulative estimation method, and scenario 2

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	3.6	80.8	1.01 (0.44)	0.89 (0.47), 1000	6.33 (2.17)	6 (3 10)
PS Matching 1:4	7.3	96.9	0.92 (0.40)	0.90 (0.41), 1000	4.56 (2.38)	4 (1 9)
PS Stratification	6.8	95.7	0.91 (0.37)	0.88 (0.38), 1000	4.25 (2.46)	4 (1 9)
DRS Matching 1:1	3.4	82.4	1.05 (0.46)	0.93 (0.50), 1000	6.15 (2.17)	6 (3 10)
DRS Matching 1:4	8.7	97.9	0.94 (0.42)	0.92 (0.43), 1000	4.50 (2.40)	4 (1 9)
DRS Stratification	8.4	97.4	0.93 (0.36)	0.92 (0.37), 1000	4.01 (2.37)	4 (1 9)
Regression	6.5	95.8	0.89 (0.35)	0.87 (0.36), 999	4.26 (2.43)	4 (1 9)

Table B7-6. Common treatment and rare outcome, cumulative estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	3.4	77.7	1.02 (0.41)	0.89 (0.45), 998	6.34 (2.25)	6 (3 10)
PS Matching 1:4	6.4	95.9	0.93 0.43	0.91 (0.43), 998	4.69 (2.46)	4 (1 9)
PS Stratification	6.4	96.3	0.91 0.36	0.89 (0.37), 998	4.37 (2.53)	4 (1 9)
DRS Matching 1:1	3.8	79.9	1.06 0.41	0.92 (0.46), 996	6.16 (2.10)	6 (3 10)
DRS Matching 1:4	7.8	95.9	0.97 0.48	0.95 (0.48), 996	4.54 (2.44)	4 (1 9)
DRS Stratification	8.2	96.8	0.93 0.36	0.91 (0.37), 996	4.11 (2.40)	4 (1 9)
Regression	6.0	96.1	0.90 0.35	0.88 (0.37), 998	4.35 (2.52)	4 (1 9)

Table B8-1. Common treatment and moderate outcome, lookwise estimation method, and scenario 1

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	4.7	99.8	0.90 (0.40)	0.90 (0.40), 1000	3.55 (1.65)	3 (1, 7)
PS Matching 1:4	6.5	100	0.81 (0.26)	0.81 (0.26), 1000	2.34 (1.25)	2 (1, 5)
PS Stratification	6.6	100	0.79 (0.23)	0.79 (0.23), 1000	2.11 (1.11)	2 (1, 4)
DRS Matching 1:1	4.4	100	0.88 (0.36)	0.88 (0.36), 1000	3.58 (1.67)	3 (1, 7)
DRS Matching 1:4	7.8	100	0.82 (0.26)	0.82 (0.26), 1000	2.33 (1.23)	2 (1, 5)
DRS Stratification	7.6	100	0.79 (0.23)	0.79 (0.23), 1000	2.11 (1.13)	2 (1 4)

Table B8-2. Common treatment and moderate outcome, lookwise estimation method, and scenario 2

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	3.8	100	0.88 (0.39)	0.88 (0.39), 1000	3.45 (1.64)	3 (1, 7)
PS Matching 1:4	9.2	100	0.82 (0.25)	0.82 (0.25), 1000	2.26 (1.19)	2 (1, 4)
PS Stratification	6.2	100	0.80 (0.23)	0.80 (0.23), 1000	2.10 (1.14)	2 (1, 4)
DRS Matching 1:1	7.2	99.9	0.91 (0.40)	0.91 (0.40), 1000	3.37 (1.64)	3 (1, 7)
DRS Matching 1:4	12.2	100	0.84 (0.26)	0.83 (0.26), 1000	2.22 (1.18)	2 (1, 4)
DRS Stratification	11.2	100	0.82 (0.23)	0.82 (0.23), 1000	1.98 (1.07)	2 (1, 4)

Table B8-3. Common treatment and moderate outcome, lookwise estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	5.3	100	0.93 (0.40)	0.93 (0.40), 1000	3.26 (1.57)	3 (1, 6)
PS Matching 1:4	8.6	100	0.85 (0.27)	0.85 (0.27), 1000	2.14 (1.11)	2 (1, 4)
PS Stratification	8.1	100	0.84 (0.23)	0.83 (0.23), 1000	1.88 (0.96)	2 (1, 4)
DRS Matching 1:1	6.4	100	0.89 (0.39)	0.89 (0.39), 1000	3.42 (1.55)	3 (1, 6)
DRS Matching 1:4	11.1	100	0.83 (0.26)	0.83 (0.26), 1000	2.19 (1.12)	2 (1, 4)
DRS Stratification	11.2	100	0.82 (0.23)	0.82 (0.23), 1000	1.94 (1.00)	2 (1, 4)

Table B8-4. Common treatment and moderate outcome, cumulative estimation method, and scenario 1

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	4.7	99.9	0.88 (0.40)	0.88 (0.40), 1000	3.64 (1.68)	3 (1, 7)
PS Matching 1:4	6.5	100	0.81 (0.26)	0.81 (0.26), 1000	2.37 (1.29)	2 (1, 5)
PS Stratification	6.6	100	0.78 (0.23)	0.78 (0.23), 1000	2.15 (1.15)	2 (1, 4)
DRS Matching 1:1	4.4	99.9	0.89 (0.39)	0.88 (0.39), 1000	3.56 (1.69)	3 (1, 7)
DRS Matching 1:4	7.8	100	0.82 (0.26)	0.82 (0.26), 1000	2.33 (1.22)	2 (1, 5)
DRS Stratification	7.6	100	0.79 (0.23)	0.79 (0.23), 1000	2.09 (1.12)	2 (1, 4)
regression	6.8	100	0.78 (0.23)	0.78 (0.23), 1000	2.15 (1.14)	2 (1, 4)

Table B8-5. Common treatment and moderate outcome, cumulative estimation method, and scenario 2

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	3.8	99.9	0.88 (0.41)	0.88 (0.41), 1000	3.53 (1.69)	3 (1, 7)
PS Matching 1:4	9.2	100	0.81 (0.26)	0.81 (0.26), 1000	2.30 (1.23)	2 (1, 5)
PS Stratification	6.2	100	0.79 (0.23)	0.79 (0.23), 1000	2.17 (1.17)	2 (1, 4)
DRS Matching 1:1	7.2	99.9	0.92 (0.40)	0.92 (0.40), 1000	3.36 (1.65)	3 (1, 7)
DRS Matching 1:4	12.2	100	0.84 (0.26)	0.84 (0.26), 1000	2.17 (1.14)	2 (1, 4)
DRS Stratification	11.2	100	0.83 (0.23)	0.83 (0.23), 1000	1.95 (1.03)	2 (1, 4)
regression	6.2	100	0.79 (0.23)	0.79 (0.23), 1000	2.18 (1.19)	2 (1, 4)

Table B8-6. Common treatment and moderate outcome, cumulative estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	5.3	100	0.86 (0.38)	0.86 (0.38), 1000	3.50 (1.64)	3 (1, 6)
PS Matching 1:4	8.6	100	0.81 (0.26)	0.81 (0.26), 1000	2.28 (1.24)	2 (1, 5)
PS Stratification	8.1	100	0.79 (0.23)	0.79 (0.23), 1000	2.05 (1.10)	2 (1, 4)
DRS Matching 1:1	6.4	100	0.90 (0.40)	0.90 (0.40), 1000	3.36 (1.58)	3 (1, 6)
DRS Matching 1:4	11.1	100	0.84 (0.27)	0.84 (0.27), 1000	2.17 (1.12)	2 (1, 4)
DRS Stratification	11.2	100	0.83 (0.23)	0.83 (0.23), 1000	1.92 (1.01)	2 (1, 4)
regression	7.9	100	0.79 (0.23)	0.79 (0.23), 1000	2.09 (1.14)	1 (1, 4)

Table B9-1. Common treatment and common outcome, lookwise estimation method, and scenario 1

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	5.8	100	0.76 (0.22)	0.76 (0.22), 1000	1.55 (0.67)	1 (1, 3)
PS Matching 1:4	5.6	100	0.70 (0.16)	0.70 (0.16), 1000	1.16 (0.37)	1 (1, 2)
PS Stratification	5.5	100	0.70 (0.15)	0.70 (0.15), 1000	1.11 (0.32)	1 (1, 2)
DRS Matching 1:1	6.3	100	0.77 (0.22)	0.77 (0.22), 1000	1.50 (0.62)	1 (1, 3)
DRS Matching 1:4	6.3	100	0.71 (0.16)	0.71 (0.16), 1000	1.15 (0.36)	1 (1, 2)
DRS Stratification	6.8	100	0.70 (0.15)	0.70 (0.15), 1000	1.10 (0.32)	1 (1, 2)

Table B9-2. Common treatment and common outcome, lookwise estimation method, and scenario 2

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	5.5	100	0.75 (0.22)	0.75 (0.22), 1000	1.50 (0.63)	1 (1, 3)
PS Matching 1:4	11.0	100	0.71 (0.16)	0.71 (0.16), 1000	1.15 (0.39)	1 (1, 2)
PS Stratification	5.0	100	0.70 (0.15)	0.70 (0.15), 1000	1.12 (0.35)	1 (1, 2)
DRS Matching 1:1	11.4	100	0.81 (0.23)	0.81 (0.23), 1000	1.38 (0.56)	1 (1, 2)
DRS Matching 1:4	17.8	100	0.76 (0.17)	0.76 (0.17), 1000	1.11 (0.33)	1 (1, 2)
DRS Stratification	18.1	100	0.74 (0.16)	0.74 (0.16), 1000	1.07 (0.27)	1 (1, 2)

Table B9-3. Common treatment and common outcome, lookwise estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	2.2	100	0.74 (0.22)	0.74 (0.22), 1000	1.57 (0.66)	1 (1, 3)
PS Matching 1:4	3.4	100	0.69 (0.16)	0.69 (0.16), 1000	1.19 (0.40)	1 (1, 2)
PS Stratification	2.9	100	0.68 (0.15)	0.68 (0.15), 1000	1.14 (0.35)	1 (1, 2)
DRS Matching 1:1	3.7	100	0.78 (0.22)	0.78 (0.22), 1000	1.47 (0.62)	1 (1, 3)
DRS Matching 1:4	5.3	100	0.71 (0.17)	0.71 (0.17), 1000	1.17 (0.39)	1 (1, 2)
DRS Stratification	4.4	100	0.71 (0.16)	0.71 (0.16), 1000	1.12 (0.32)	1 (1, 2)

Table B9-4. Common treatment and common outcome, cumulative estimation method, and scenario 1

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	5.4	100	0.75 (0.22)	0.75 (0.22), 1000	1.56 (0.69)	1 (1, 3)
PS Matching 1:4	5.9	100	0.70 (0.16)	0.70 (0.16), 1000	1.16 (0.37)	1 (1,2)
PS Stratification	5.1	100	0.70 (0.15)	0.70 (0.15), 1000	1.11 (0.32)	1 (1,2)
DRS Matching 1:1	5.7	100	0.77 (0.22)	0.77 (0.22), 1000	1.50 (0.62)	1 (1,2)
DRS Matching 1:4	6.5	100	0.71 (0.16)	0.71 (0.16), 1000	1.15 (0.37)	1 (1,2)
DRS Stratification	6.8	100	0.70 (0.15)	0.70 (0.15), 1000	1.10 (0.30)	1 (1,2)
Regression	5.0	100	0.70 (0.15)	0.70 (0.15), 1000	1.11 (0.32)	1 (1,2)

Table B9-5. Common treatment and common outcome, cumulative estimation method, and scenario 2

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	4.7	100	0.75 (0.21)	0.75 (0.21), 1000	1.49 (0.64)	1 (1,3)
PS Matching 1:4	10.3	100	0.71 (0.15)	0.71 (0.15), 1000	1.14 (0.37)	1 (1,2)
PS Stratification	6.0	100	0.70 (0.15)	0.70 (0.15), 1000	1.11 (0.33)	1 (1,2)
DRS Matching 1:1	11.2	100	0.81 (0.23)	0.81 (0.23), 1000	1.40 (0.58)	1 (1,2)
DRS Matching 1:4	17.0	100	0.75 (0.16)	0.75 (0.16), 1000	1.11 (0.32)	1 (1,2)
DRS Stratification	15.3	100	0.74 (0.16)	0.74 (0.16), 1000	1.07 (0.25)	1 (1,2)
Regression	5.2	100	0.70 (0.15)	0.70 (0.15), 1000	1.12 (0.34)	1 (1,2)

Table B9-6. Common treatment and common outcome, cumulative estimation method, and scenario 3

		TREATMENT PARAMETER θ_Z=0.69 (OR=2)				
Design and Method	**θ_Z=0 (OR=1) Type 1 Error %**	**Signaled %**	**MEAN $\hat{\theta}_Z$ (STD)**		**Mean Time to Signal Detection (STD)**	**Median Time to Signal Detection (5%, 95%)**
			Signaled	**Signaled and Not, N**		
PS Matching 1:1	5.7	100	0.75 (0.22)	0.75 (0.22), 1000	1.56 (0.69)	1 (1,3)
PS Matching 1:4	8.3	100	0.70 (0.16)	0.70 (0.16), 1000	1.16 (0.38)	1 (1,2)
PS Stratification	6.6	100	0.70 (0.15)	0.70 (0.15), 1000	1.11 (0.32)	1 (1,2)
DRS Matching 1:1	8.0	100	0.78 (0.22)	0.78 (0.22), 1000	1.47 (0.62)	1 (1,3)
DRS Matching 1:4	10.9	100	0.72 (0.16)	0.72 (0.16), 1000	1.13 (0.35)	1 (1,2)
DRS Stratification	10.8	100	0.72 (0.15)	0.72 (0.15), 1000	1.09 (0.29)	1 (1,2)
Regression	6.0	100	0.70 (0.15)	0.70 (0.15), 1000	1.11 (0.33)	1 (1,2)

IX. APPENDIX C

Table C1. Incident GI bleed within six months and type of NSAID exposures by year and site

SITE	YEAR OF DRUG INITIATION	GI BLEED WITHIN 6 MONTHS		TYPE OF NSAID DRUG	
		No N (%)	Yes N (%)	Non-selective NSAID N (%)	COX-2 NSAID N (%)
1	2008	330014 (99.85%)	484 (0.15%)	315716 (95.53%)	14782 (4.47%)
	2009	542085 (99.86%)	756 (0.14%)	524890 (96.69%)	17951 (3.31%)
	2010	427690 (99.87%)	548 (0.13%)	416326 (97.22%)	11912 (2.78%)
	2011	366124 (99.87%)	479 (0.13%)	358455 (97.78%)	8148 (2.22%)
2	2008	193535 (99.80%)	392 (0.20%)	183781 (94.77%)	10146 (5.23%)
	2009	162160 (99.79%)	340 (0.21%)	156018 (96.01%)	6482 (3.99%)
	2010	144572 (99.79%)	300 (0.21%)	139656 (96.40%)	5216 (3.60%)
	2011	139742 (99.80%)	286 (0.20%)	135141 (96.51%)	4887 (3.49%)
15	2008	107165 (99.95%)	53 (0.05%)	106998 (99.79%)	220 (0.21%)
	2009	96063 (99.95%)	47 (0.05%)	95870 (99.75%)	240 (0.25%)
	2010	88133 (99.95%)	41 (0.05%)	87939 (99.73%)	235 (0.27%)
	2011	87914 (99.95%)	42 (0.05%)	87640 (99.64%)	316 (0.36%)

Table C2. Incident GI bleed outcome: results for Propensity Score matched samples (4:1) (4 Nonselective NSAIDs matched to 1 COX-2) ; Full cohort with matches not restricted to caliper of +0.05

Look	Time[1]	Nonselective NSAID		COX-2 NSAID		HR[4] Nonselective NSAID vs COX-2	Sequential analysis test statistic[5]	Test statistic boundary[6]	Signal[7]
		GI bleeds $_{cum}$[2]	Person years$_{cum}$[3]	GI bleeds $_{cum}$	Person years$_{cum}$				
Lookwise PS[8]									
1	2008	287	45849	63	11510	1.15	0.968	2.1040	No
2	2009	603	90831	130	22824	1.17	1.613	2.0323	No
3	2010	824	122314	174	30703	1.19	2.116	2.0514	Yes
4	2011	995	146961	224	36878	1.12	1.511	2.0508	
Cumulative PS[9]									
1	2008	294	45864	63	11510	1.18	1.199	2.1040	No
2	2009	600	90865	130	22824	1.16	1.552	2.0323	No
3	2010	804	122337	174	30703	1.17	1.812	2.0514	No
4	2011	976	147016	224	36878	1.09	1.178	2.0508	No

[1]Timing of each look; Later looks include information from the preceding time intervals

[2]GI bleeds$_{cum}$ is the total number of incident GI bleed events up to and including the year of each look (within each selected drug group)

[3]Person years$_{cum}$ is the total accumulated exposure time up to and including the year of each look (within each selected drug group)

[4]HR is the hazard ratio for nonselective NSAID vs COX-2 drugs for all times up to and including each look

[5]Test statistic is the drug comparison parameter estimate / standard error from the proportional hazards model

[6]Test statistic boundary: sequential analysis boundary estimate

[7]Signal indicates when the test statistic first exceeds the test statistic boundary

[8]Lookwise estimation used single year data when estimating scores

[9]Cumulative estimation used all years up to and including the current year when estimating scores

Table C3. Incident GI bleed outcome: results for Disease Risk Score matched samples (4:1) (4 Nonselective NSAIDs matched to 1 COX-2) ; Full cohort with matches not restricted to caliper of +0.05

Look	Time[1]	Nonselective NSAID		COX-2 NSAID		HR[4] Nonselective NSAID vs COX-2	Sequential analysis test statistic[5]	Test statistic boundary[6]	Signal[7]
		GI bleeds $_{cum}$[2]	Person years$_{cum}$[3]	GI bleeds$_{cum}$	Person years$_{cum}$				
Lookwise DRS[8]									
1	2008	318	45765	63	11510	1.32	1.972	2.1040	No
2	2009	666	90704	130	22824	1.32	2.857	2.0323	Yes
3	2010	892	122154	174	30703	1.32	3.323	2.0514	
4	2011	1061	146843	224	36878	1.22	2.647	2.0508	
Cumulative DRS[9]									
1	2008	311	45772	63	11510	1.29	1.820	2.1040	No
2	2009	651	90726	130	22824	1.31	2.759	2.0323	Yes
3	2010	848	122205	174	30703	1.27	2.821	2.0514	
4	2011	1017	146876	224	36878	1.18	2.194	2.0508	

[1]Timing of each look; Later looks include information from the preceding time intervals

[2]GI bleeds$_{cum}$ is the total number of incident GI bleed events up to and including the year of each look (within each selected drug group)

[3]Person years$_{cum}$ is the total accumulated exposure time up to and including the year of each look (within each selected drug group)

[4]HR is the hazard ratio for nonselective NSAID vs COX-2 drugs for all times up to and including each look

[5]Test Statistic is the drug comparison parameter estimate / standard error from the proportional hazards model

[6]Test statistic boundary: sequential analysis boundary estimate

[7]Signal indicates when the test statistic first exceeds the test statistic boundary

[8]Lookwise estimation used single year data when estimating scores

[9]Cumulative estimation used all years up to and including the current year when estimating scores

Table C4. Incident GI bleed outcome: results for Propensity matched samples (1:1) (1 Nonselective NSAIDs matched to 1 COX-2)

Look	Time[1]	Nonselective NSAID		COX-2 NSAID		HR[4] Nonselective NSAID vs COX-2	Sequential analysis test statistic[5]	Test statistic boundary[6]	Signal[7]
		GI bleeds $_{cum}$[2]	Person years$_{cum}$[3]	GI bleeds$_{cum}$	Person years$_{cum}$				
Lookwise PS[8]									
1	2008	70	11451	63	11510	1.24	1.170	2.1040	No
2	2009	134	22683	130	22824	1.09	0.699	2.0323	No
3	2010	179	30563	174	30701	1.08	0.658	2.0514	No
4	2011	223	36712	224	36877	1.04	0.389	2.0508	No
Cumulative PS[9]									
1	2008	66	11459	63	11510	1.20	1.007	2.1040	No
2	2009	134	22689	130	22824	1.07	0.505	2.0323	No
3	2010	178	30563	174	30701	1.06	0.495	2.0514	No
4	2011	230	36728	224	36877	1.04	0.387	2.0508	No

[1]Timing of each look; Later looks include information from the preceding time intervals

[2]GI bleeds$_{cum}$ is the total number of incident GI bleed events up to and including the year of each look (within each selected drug group)

[3]Person years$_{cum}$ is the total accumulated exposure time up to and including the year of each look (within each selected drug group)

[4]HR is the hazard ratio for nonselective NSAID vs COX-2 drugs for all times up to and including each look

[5]Test statistic is the drug comparison parameter estimate / standard error from the proportional hazards model

[6]Test statistic boundary: sequential analysis boundary estimate

[7]Signal indicates when the test statistic first exceeds the test statistic boundary

[8]Lookwise estimation used single year data when estimating scores

[9]Cumulative estimation used all years up to and including the current year when estimating scores

Table C5. Incident GI bleed outcome: results for Disease Risk Score matched samples (1:1) (1 Nonselective NSAIDs matched to 1 COX-2)

Look	Time[1]	Nonselective NSAID		COX-2 NSAID		HR[4] Nonselective NSAID vs COX-2	Sequential analysis test statistic[5]	Test statistic boundary[6]	Signal[7]
		GI bleeds $_{cum}$[2]	Person years$_{cum}$[3]	GI bleeds $_{cum}$	Person years$_{cum}$				
Lookwise DRS[8]									
1	2008	82	11349	63	11403	1.51	2.325	2.1040	Yes
2	2009	163	22426	130	22602	1.34	2.399	2.0323	
3	2010	208	30184	174	30366	1.24	2.054	2.0514	
4	2011	254	36191	224	36388	1.18	1.744	2.0508	
Cumulative DRS[9]									
1	2008	86	11341	63	11403	1.51	2.366	2.1040	Yes
2	2009	169	22437	130	22602	1.31	2.781	2.0323	
3	2010	212	30193	174	30366	1.27	2.612	2.0514	
4	2011	252	36208	224	36388	1.18	2.075	2.0508	

[1]Timing of each look; Later looks include information from the preceding time intervals

[2]GI bleeds$_{cum}$ is the total number of incident GI bleed events up to and including the year of each look (within each selected drug group) , only two sites included due to DRS data issues at third site

[3]Person years$_{cum}$ is the total accumulated exposure time up to and including the year of each look (within each selected drug group), only two sites included due to DRS data issues at third site

[4]HR is the hazard ratio for nonselective NSAID vs COX-2 drugs for all times up to and including each look

[5]Test Statistic is the drug comparison parameter estimate / standard error from the proportional hazards model

[6]Test statistic boundary: sequential analysis boundary estimate

[7]Signal indicates when the test statistic first exceeds the test statistic boundary

[8]Lookwise estimation used single year data when estimating scores

[9]Cumulative estimation used all years up to and including the current year when estimating scores