PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2012; **21**(S2): 69–80
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/pds.3263

ORIGINAL REPORT

One-to-many propensity score matching in cohort studies

Jeremy A. Rassen[1]*, Abhi A. Shelat[2], Jessica Myers[1], Robert J. Glynn[1], Kenneth J. Rothman[3] and Sebastian Schneeweiss[1]

[1] *Division of Pharmacoepidemiology and Pharmacoeconomics; Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA*
[2] *Department of Computer Science, University of Virginia, Charlottesville, VA, USA*
[3] *RTI International, Research Triangle Park, NC, USA*

ABSTRACT

Background Among the large number of cohort studies that employ propensity score matching, most match patients 1:1. Increasing the matching ratio is thought to improve precision but may come with a trade-off with respect to bias.
Objective To evaluate several methods of propensity score matching in cohort studies through simulation and empirical analyses.
Methods We simulated cohorts of 20 000 patients with exposure prevalence of 10%–50%. We simulated five dichotomous and five continuous confounders. We estimated propensity scores and matched using digit-based greedy ("greedy"), pairwise nearest neighbor within a caliper ("nearest neighbor"), and a nearest neighbor approach that sought to balance the scores of the comparison patient above and below that of the treated patient ("balanced nearest neighbor"). We matched at both fixed and variable matching ratios and also evaluated sequential and parallel schemes for the order of formation of 1:*n* match groups. We then applied this same approach to two cohorts of patients drawn from administrative claims data.
Results Increasing the match ratio beyond 1:1 generally resulted in somewhat higher bias. It also resulted in lower variance with variable ratio matching but higher variance with fixed. The parallel approach generally resulted in higher mean squared error but lower bias than the sequential approach. Variable ratio, parallel, balanced nearest neighbor matching generally yielded the lowest bias and mean squared error.
Conclusions 1:*n* matching can be used to increase precision in cohort studies. We recommend a variable ratio, parallel, balanced 1:*n*, nearest neighbor approach that increases precision over 1:1 matching at a small cost in bias. Copyright © 2012 John Wiley & Sons, Ltd.

KEY WORDS—propensity scores; confounding factors (epidemiology); epidemiologic methods; comparative effectiveness research

Received 19 August 2011; Revised 23 February 2012; Accepted 24 February 2012

INTRODUCTION

Epidemiologists have long employed matching in cohort studies, and matched cohort studies may be particularly applicable in automated safety surveillance systems and other scenarios. Whereas matching was traditionally performed on specific factors—age, sex, days on treatment[1]—today's matching is often carried out on a summary score, such as a propensity score[2–4] or disease risk score.[5] In cohort studies, matching on propensity scores offers investigators the ability to balance treatment groups across all putative risk factors, and allows easy inspection of the achieved balance across measured covariates. It excludes those subjects in the non-overlapping ranges of the score, thereby giving an estimate of the treatment effect among the treated, an important clinical measure.[6–8] The matching process serves a function similar to propensity score trimming and improves the validity of the estimate.[9] 1:1 matching on propensity scores is often performed using SAS-based greedy matching algorithm,[10] which offers a fast way to get approximately nearest neighbor matches.[10,11] Nearest neighbor matching, although shown to provide better balance among treatment groups, is not frequently used in epidemiology.[12]

Cohort study matching at ratios of 1:*n*, with either a fixed or variable *n*, can yield higher precision and thus smaller confidence intervals than does simple 1:1 matching. It has also been known to increase bias, because second matches will generally be of lower quality than the first.[13] When going beyond 1:1 matching

*Correspondence to: J. A. Rassen, Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 1620 Tremont Street Suite 3030, Boston, MA 02120, USA. E-mail: jrassen@post.harvard.edu
Dr. Rassen is a recipient of a career development award from Agency for Healthcare Research and Quality (K01 HS018088). The Division of Pharmacoepidemiology received gifts from IBM Netezza and Tableau Software.

ratios, Ming and Rosenbaum recommend employing variable ratio techniques, which reduce bias as compared with fixed ratio matching but also result in a loss in transparency of a "Table 1" presentation of covariate balance in the patient cohort.[14] Variable ratio matching retains more exposed subjects than fixed ratio by not dropping those without the set number of comparison group matches.

1:*n* matching without replacement is frequently accomplished by creating a 1:1 matched cohort and then adding second-level, third-level, and higher-level matches from among the remaining patients.[15,16] We refer to this as a sequential, "everyone gets firsts before anyone gets seconds" approach. The advantage of this approach is that no treated patient is "starved" of his or her single best available comparison patient match as a result of using that best match in a secondary position for another treated patient. At the same time, the sequential approach may lessen the quality of certain matched sets by potentially downgrading the quality of a treated patient's secondary matches. If enough matched sets are affected, then the distance between the treated and comparison groups in the overall cohort may be larger than necessary, resulting in a biased point estimate.

Here, we examine alternatives to the Parsons greedy matching methodology, including a true nearest neighbor approach that minimizes within-set distances. We also examine several schemes for matching that yield 1:*n* cohorts. We examine the performance of these schemes through simulation and empirical studies, as applied in the context of cohort studies of drug effects in healthcare databases.

METHODS

Treatment and comparison groups

Throughout this paper, we refer to the two exposure categories as the treatment and the comparison groups and assume that a single treated patient is matched to one or more comparison patients.

In observational research, the goal of matching is to create treatment and comparison groups that are balanced on all measured confounders. Matching on a balancing score will yield, in expectation, balance between treatment groups for the covariates included in the score.[3] Although it is a common practice to match on a propensity score,[17] it is also possible to match on a summary disease risk score, dichotomous variables (e.g., sex), or continuous values (e.g., logit of propensity score, age).

Type of matching and terminology

Although greedy matching has a general meaning in the biostatistics literature, the term in epidemiology tends to refer to the SAS-based implementation of greedy matching by Parsons.[10,18] Parsons' approach matches patients on decreasing levels of precision of the propensity score. Treated patients are considered sequentially.[10] Each treated patient is matched to a comparison patient whose score equals that of the treated patient to at least the fifth digit. When all matches at the fifth digit are exhausted, the process begins again at the fourth digit and so forth.

This approach to greedy matching is an efficient approximation of a type of nearest neighbor matching, in which each treated patient is matched to the unmatched comparison patient with the closest propensity score, with "closest" commonly defined as the difference in the two patients' scores. A maximum allowable distance (the "caliper") is often imposed.

The use of this type of nearest neighbor matching has been in part limited by the lack of efficient software to compute the best matches; to our knowledge, existing software either computes all possible pairings of treated and comparison patients and selects the nearest pairings within a predefined caliper[19] or finds treated patients' best comparison patient matches on the basis of a single ordering (high score to low, low score to high, random, order appearing in the data).[16] As an alternative, we have implemented a nearest neighbor technique that guarantees computation of the best matches, gives consistent results independent of any ordering of patients, and avoids the exponential scaling of required time and memory with the number of subjects. In typical configurations, it executes in less than 1 second (see Appendix A).

Our pairwise approach to nearest neighbor matching yields a cohort in which the distance between each pair of patients is minimized, but the overall distance between the treated and comparison groups may not be optimal. In practice, we believe the difference between pairwise nearest neighbor and optimal nearest neighbor matching is minimal, and pairwise nearest neighbor matching is far faster to compute. Appendix B demonstrates a case in which the results of pairwise nearest neighbor and optimal nearest neighbor matching will differ. Because of what we perceive to be small differences in the amount of confounding adjustment offered by the two techniques, and the substantially greater compute time required by optimal matching, we consider only pairwise nearest neighbor matching in this paper.

Unfortunately, there has been some inconsistency in matching terminology in the epidemiology and biostatistics literature. In this paper, we refer to pairwise nearest neighbor matching within a fixed caliper simply as nearest neighbor matching. Other literature refers to this approach as greedy matching with a caliper and refers to what we describe as optimal nearest neighbor

DOI: 10.1002/pds

Table 1. Matching simulation results for base exposure prevalence of 30%

Matching scheme	Standardized differences of measured variables*				Mean standardized distance	Max. standard distance	Mean number of matched sets	Mean % of treated patients matched	Mean matching ratio (1:*n*)	Mean treatment effect (SD)**	Mean bias (%)	Mean squared error
	C_1	C_5	D_1	D_5								
Unmatched	0.199	0.577	0.209	0.560	0.391	0.407	6001	30.00	2.3	10.43 (0.31)	−943.3	90.678
1:1 Matching												
Nearest neighbor matching	0.000	0.000	0.000	0.000	0.000	0.005	4272	50.00	1.0	0.99 (0.40)	0.7	0.163
Digit-based greedy matching	0.000	0.001	0.000	0.001	0.000	0.004	4285	50.00	1.0	1.01 (0.41)	−0.6	0.166
2:1 Matching												
Nearest neighbor matching												
Sequential variable ratio	0.029	0.092	0.031	0.089	0.061	0.065	4271	39.65	1.5	1.01 (0.38)	−0.9	0.146
Parallel variable ratio	0.017	0.054	0.019	0.048	0.036	0.041	3683	36.19	1.8	1.01 (0.39)	−1.1	0.154
Sequential fixed ratio	0.000	0.001	0.000	0.001	0.000	0.007	2230	33.33	2.0	1.01(0.50)	10.9	0.250
Parallel fixed ratio	0.000	0.001	0.000	0.000	0.000	0.005	2819	33.33	2.0	0.99 (0.43)	0.5	0.188
Balanced nearest neighbor matching												
Sequential variable ratio	0.028	0.090	0.030	0.089	0.060	0.064	4271	40.44	1.5	1.01 (0.38)	11.1	0.147
Parallel variable ratio	0.021	0.067	0.023	0.066	0.045	0.050	3973	39.43	1.5	1.01 (0.39)	11.0	0.151
Sequential fixed ratio	0.001	0.001	0.000	0.000	0.000	0.008	2020	33.33	2.0	1.01 (0.52)	11.2	0.274
Parallel fixed ratio	0.000	0.001	0.000	0.000	0.000	0.007	2133	33.33	2.0	1.00 (0.50)	0.1	0.250
Digit-based greedy matching												
Sequential variable ratio	0.029	0.093	0.031	0.091	0.062	0.066	4284	39.76	1.5	1.02 (0.38)	12.2	0.146
Parallel variable ratio	0.013	0.040	0.014	0.036	0.026	0.031	3882	40.04	1.5	1.01 (0.40)	11.4	0.159
Sequential fixed ratio	0.000	0.001	0.000	0.001	0.000	0.007	2205	33.33	2.0	1.02 (0.50)	12.0	0.252
Parallel fixed ratio	0.000	0.001	0.000	0.002	0.000	0.008	1935	33.33	2.0	1.00 (0.54)	0.2	0.297
3:1 Matching												
Nearest neighbor matching												
Sequential variable ratio	0.046	0.146	0.049	0.147	0.097	0.103	4271	35.03	1.9	1.01 (0.37)	10.7	0.137
Parallel variable ratio	0.032	0.102	0.036	0.096	0.068	0.074	3463	30.42	2.3	1.01 (0.39)	10.9	0.154
Sequential fixed ratio	0.000	0.001	10.001	0.001	0.000	0.010	1423	25.00	3.0	1.00 (0.60)	0.5	0.355
Parallel fixed ratio	0.000	0.000	0.000	0.000	0.000	0.007	2007	25.00	3.0	1.02(0.50)	11.8	0.249
Balanced nearest neighbor matching												
Sequential variable ratio	0.046	0.145	0.049	0.147	0.097	0.102	4271	35.19	1.8	1.01 (0.37)	10.6	0.138
Parallel variable ratio	0.036	0.114	0.039	0.113	0.076	0.082	3809	33.64	2.0	1.01 (0.39)	10.8	0.151
Sequential fixed ratio	0.000	0.001	0.000	0.000	0.000	0.010	1497	25.00	3.0	0.99 (0.58)	0.6	0.343
Parallel fixed ratio	10.001	0.000	0.000	0.000	0.000	0.009	1743	25.00	3.0	1.02 (0.53)	11.8	0.287
Digit-based greedy matching												
Sequential variable ratio	0.046	0.147	0.049	0.148	0.098	0.103	4285	35.19	1.8	1.01 (0.37)	11.5	0.138
Parallel variable ratio	0.022	0.069	0.024	0.064	0.045	0.052	3731	36.01	1.8	1.01 (0.40)	10.8	0.163
Sequential fixed ratio	0.000	0.001	0.000	0.001	0.000	0.009	1402	25.00	3.0	1.00 (0.60)	0.3	0.362
Parallel fixed ratio	0.000	0.000	0.000	0.001	0.001	0.012	1114	25.00	3.0	1.00 (0.66)	0.1	0.443
4:1 Matching												
Nearest neighbor matching												
Sequential variable ratio	0.058	0.183	0.061	0.189	0.123	0.129	4268	32.35	2.1	1.01 (0.36)	10.7	0.127
Parallel variable ratio	0.044	0.139	0.049	0.136	0.093	0.099	3362	27.37	2.7	1.00 (0.40)	0.2	0.161
Sequential fixed ratio	0.001	0.001	0.001	0.002	0.001	0.012	1009	20.00	4.0	0.99 (0.67)	0.5	0.450
Parallel fixed ratio	0.000	0.000	0.001	0.000	0.001	0.009	1516	20.00	4.0	1.01 (0.55)	11.5	0.301

(*Continues*)

DOI: 10.1002/pds

Table 1. (Continued)

Matching scheme	Standardized differences of measured variables*										
	C_1	C_5	D_1	D_5	Mean standardized distance	Max. standard distance	Mean number of matched sets	Mean % of treated patients matched	Mean matching ratio (1:*n*)	Mean treatment effect (SD)**	Mean bias (%)	Mean squared error
Balanced nearest neighbor matching												
Sequential variable ratio	0.057	0.181	0.060	0.187	0.121	0.127	4268	32.64	2.1	1.00 (0.36)	10.4	0.129
Parallel variable ratio	0.045	0.141	0.048	0.142	0.094	0.102	3771	31.52	2.2	0.99 (0.40)	0.6	0.157
Sequential fixed ratio	0.000	0.001	0.001	0.001	0.001	0.014	916	20.00	4.0	0.99 (0.70)	1.2	0.494
Parallel fixed ratio	0.000	10.001	0.000	0.000	0.000	0.014	894	20.00	4.0	1.01 (0.70)	10.6	0.490
Digit-based greedy matching												
Sequential variable ratio	0.059	0.187	0.062	0.193	0.125	0.131	4283	32.18	2.1	1.03 (0.36)	13.4	0.128
Parallel variable ratio	0.028	0.088	0.031	0.085	0.059	0.067	3676	33.73	2.0	1.00 (0.41)	0.3	0.166
Sequential fixed ratio	0.002	0.005	0.001	0.007	0.003	0.015	1140	20.00	4.0	1.06 (0.62)	16.2	0.390
Parallel fixed ratio	0.000	0.001	0.000	0.002	0.001	0.016	747	20.00	4.0	1.05 (0.80)	14.5	0.647

*The variable *C1* is the first continuous variable, whereas *C5* is the fifth. The variable *D1* is the first dichotomous variable, whereas *D5* is the fifth.
**The expected treatment effect is 1.0. Any change from 1.0 is bias. SD is standard deviation.

matching as optimal matching.[6] We refer to Parsons' commonly used digit-based greedy matching approach as greedy matching to invoke the standard term in the epidemiology literature.

1:n matching

We examined a series of strategies for 1:*n* propensity score matching, which has a smaller body of literature[13,14] than does 1:1 matching.[12] In each case, we considered both fixed ratio matching, in which sets must have one treated patient and exactly *n* comparison patients, as well as variable ratio matching, in which one treated patient is matched to up to *n* comparison patients.[1,14] In a cohort study, the analysis can ignore the matching set if fixed ratio matching is applied—though at a possible cost of precision[1,20]—but variable sizes of match groups require accounting for the match using stratification by number of matches or by matched set.[13]

For each method, we considered both sequential and parallel matched set building. In sequential matched set building, we created an initial group of 1:1 matches. Then, we added second matches to the 1:1 matches, then third matches to the 2:1 matches, and so forth, with additional comparison patients added from among those who had not been previously matched. This method yielded a cohort in which the first match in each matched set was the best possible match, and each succeeding match was of equal or lesser quality. Any ties were broken randomly. The advantage of this approach is that each treated patient has the opportunity to be matched with his or her best available comparison group patient, without the potential best comparison group patient being used as a secondary match for another treated patient. However, this approach may compromise balance.

In parallel matched set building, we sought to minimize the within-set distance among the overall matched cohort. In this method, the best treated-to-comparison match is made first. Then, if the next best match would involve an already-matched treated patient, we made a second (third, up to *n*th) match for that treated patient even if there were other treated patients who had not yet been assigned a first match. Although this method should yield well-matched sets, treated patients may be "starved" of their best first match in favor of a second-position match for another treated patient.

For each combination of fixed and variable ratios, and sequential and parallel approaches, we applied the following match techniques. In all cases, we worked on the natural scale of propensity scores rather than a logit or other transformation.[12]

(1) Digit-based greedy matching. We applied a fifth digit to first digit (5 → 1) greedy matching technique

DOI: 10.1002/pds

as described previously, [10,12,15] with several modifications to the Parsons' algorithm. We (i) matched comparison patients to treated patients, with the comparison sorted by increasing propensity score and the treated patients sorted randomly; (ii) broke any ties by using the smallest match distance among possible matches; (iii) broke any remaining ties by using a random comparison patient; and (iv) substantially improved the speed of the algorithm by using advanced data structures.[21]

(2) Pairwise nearest neighbor matching. We implemented a nearest neighbor matching algorithm that minimized distance within matched sets and applied a caliper of 0.05 on the propensity score scale. Whereas others have suggested smaller calipers,[12,3] we used 0.05 to allow for ready comparison to $5 \rightarrow 1$ greedy matching. Using a smaller caliper may improve match quality but may also limit matches and thus lower precision. We believe that results at a caliper of 0.05 will overestimate any bias as compared with smaller calipers.

(3) Balanced pairwise nearest neighbor matching. Balanced nearest neighbor matching extends nearest neighbor matching by requiring that comparison patients alternate between having scores greater than (to the right of) and less than (to the left of) their matched treatment patient. Any odd-numbered match (first, third, . . .) can occur on the left or the right of the treated patient; even-numbered matches must then occur on the side opposite of where the prior odd-numbered match occurred. We implemented this extension to avoid the potential problem of comparison patients' consistently clustering on one side of the matched treated patient.

SIMULATION STUDY

We tested these approaches in a simulation study. In each run of the simulation, we created 20 000 patients. Following the design described by Austin,[12] we assigned each patient's exposure by using a binomial distribution and a base exposure prevalence from

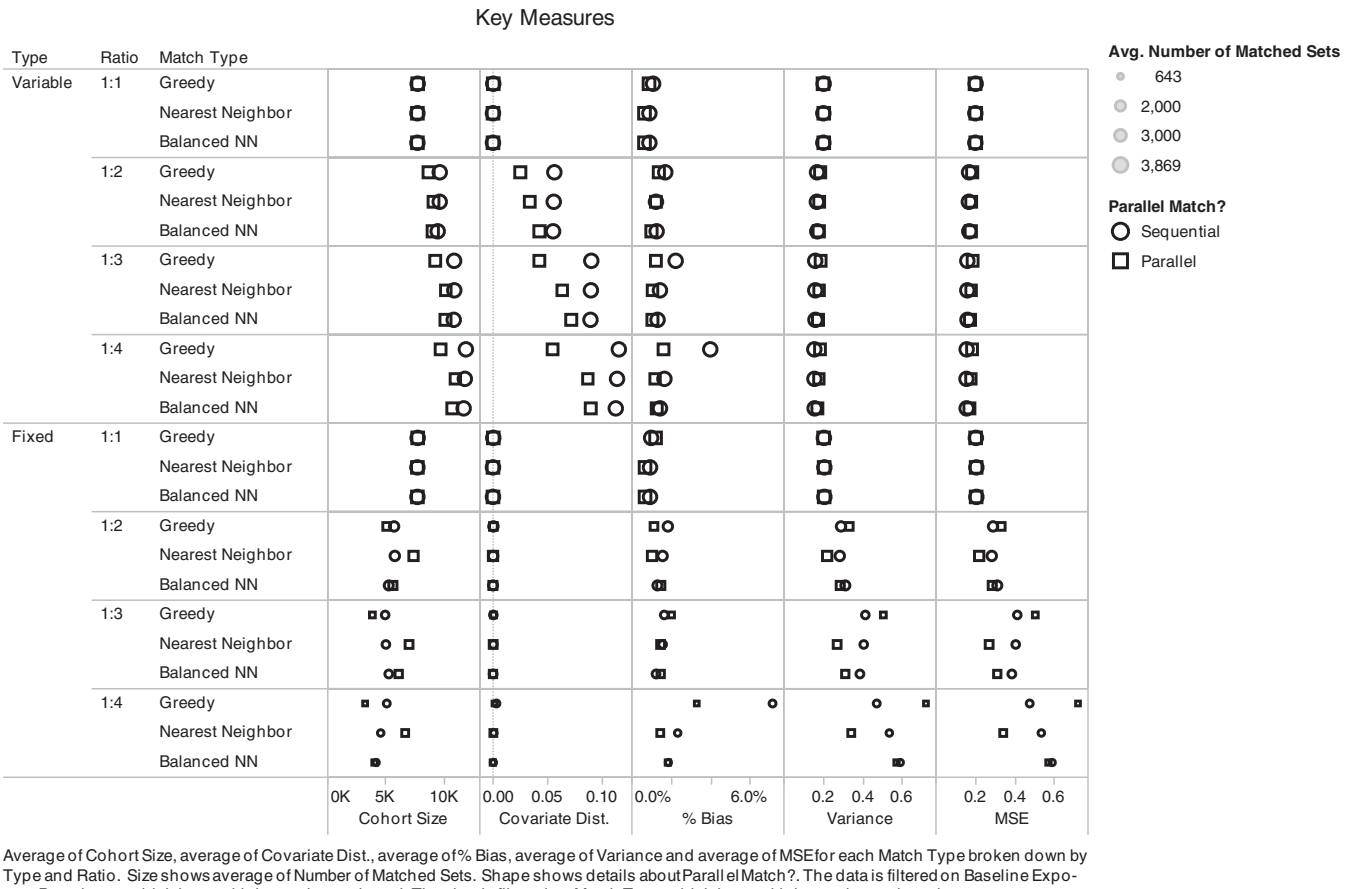

Figure 1. Observed results from simulations of various 1:*n* matching approaches, averaged over all simulation runs. Points are sized in proportion to the average size of the matched cohorts (smallest = 643; largest = 3869); circular points indicate sequential matching, whereas square points indicate parallel. NN, nearest neighbor

DOI: 10.1002/pds

10% to 50%. We created five continuous covariates and five dichotomous covariates. The first continuous covariate had a standardized difference among the exposed and unexposed of 10%, the second 20%, and so forth. The five dichotomous covariates were constructed with a baseline prevalence of 10%–50% among the unexposed and a standardized difference in prevalence between the exposed and unexposed of 10%. We simulated a continuous outcome as a function of the 10 covariates, plus a treatment effect of 1.0. We measured the standard deviation of the outcome and added to it an error that was normally distributed with mean 0 and twice the observed standard deviation. This process yielded an r^2 of approximately 0.20 when regressing the outcome as a function of the treatment and covariates.

We simulated 1000 datasets at each level of exposure prevalence. In each dataset, we matched using each of the three schemes described earlier, at ratios of 1:1 to 1:4, with fixed and variable matching ratios, and sequential and parallel approaches. To judge the quality of the matched sets, we measured the standardized distance among the 10 covariates in the matched datasets, as well as the observed treatment effect estimate and variance, and the associated mean bias and mean squared error (MSE). Our primary ("within-set") treatment effect estimates accounted for the matching and were obtained by calculating the difference within each matched set between the treated patient's outcome and the mean outcome among the comparison patients; the overall treatment effect in each simulation run was then the mean of those differences. As a secondary approach, we employed an "across sets" ordinary least squares regression approach with separate models for each number of comparison patients in the matched sets (i.e., one model for patients matched 1:1, one for patients matched 1:2, and so forth). In this case, each simulation run's overall treatment effect was estimated as the inverse-variance weighted average of the stratum-specific estimates.

EMPIRICAL STUDIES

We also applied the matching schemes to two empirical datasets, both drawn from databases of insurance claims for prescription drugs, medical services, and hospitalizations. We performed a study of initiation of nonsteroidal anti-inflammatory drug (NSAID) therapy and its effect on severe gastrointestinal (GI) complications.[22] A dichotomous exposure variable indicated a class of NSAID; non-selective NSAIDs (ibuprofen, naproxen, and diclofenac) were the comparison category,

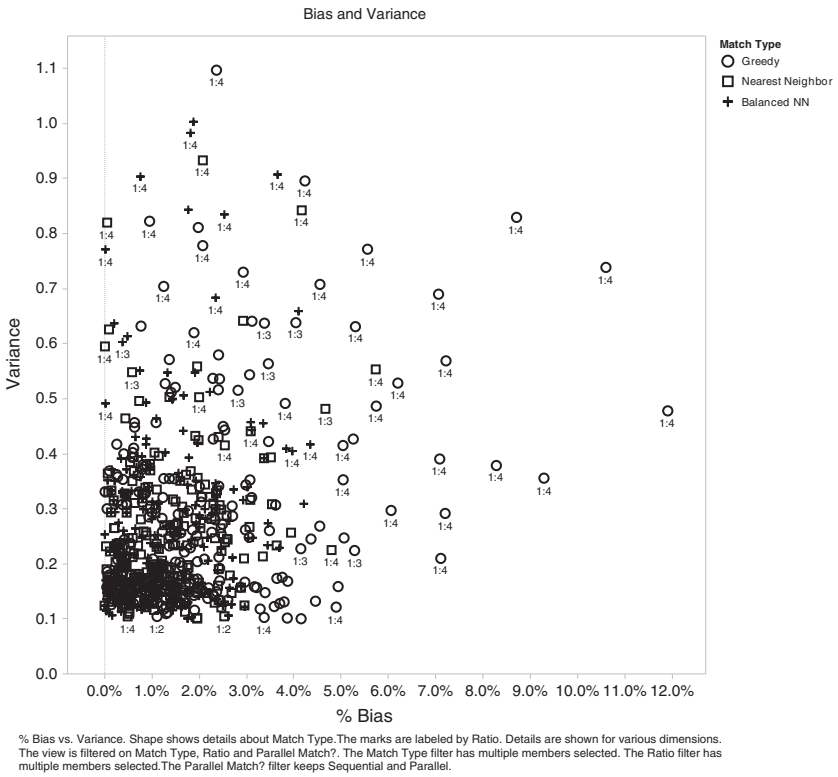

Figure 2. Bias and variance of all observed matching scenarios, averaged over all observations in that scenario. Circular points indicate greedy matching; square points indicate nearest neighbor; and pluses indicate balanced nearest neighbor

DOI: 10.1002/pds

compared with cyclo-oxygenase 2 inhibitors (coxibs; celecoxib, rofecoxib, valdecoxib) as the treatment category. We defined outcome as the cumulative risk of a GI complication (hospitalization for GI hemorrhage or peptic ulcer disease or claim for associated services) within 180 days of treatment initiation. The study was performed in a cohort drawn from Pennsylvania's Pharmaceutical Assistance Contract for the Elderly (PACE), a drug assistance program for the state's lower-income seniors. We received claims from 1994 to 2003 for those Pennsylvania's PACE participants also enrolled in Medicare. The full study design has been described in other work.[23–25] Because the treatment group was larger than the comparison group—a challenging but important case—we considered only 1:1, 1:2, and 1:3 matching ratios. Performing 2:1 or 3:1 matching, with multiple coxib patients per NSAID patient, would also have been possible.

We also considered patients who initiated cholesterol-control therapy using statin drugs alone or a statin initiated concurrently with ezetimibe (or use of the combination product Vytorin) and the subsequent risk of myocardial infarction, stroke, or death within 180 days of treatment initiation. The study was performed in a cohort drawn from pharmacy benefits manager Caremark with linked data from Medicare Parts A and B for patients 65 years and older who initiated therapy from 2005 through 2008. In this study, we considered matching at levels from 1:1 to 1:5.

To account for the variable number of patients in each matched set, we estimated both risk differences as in the simulation study as well as odds ratios by using conditional logistic regression stratified by matched set. All analyses were carried out in R version 2.13, although they could also have been performed in SAS or STATA. All programs and source code for the matching algorithms described are available as part of the Pharmacoepidemiology Toolbox,[26] available at http://www.hdpharmacoepi.org. Simulation results were analyzed on a Netezza data warehouse appliance (IBM Netezza, Marlborough, MA), and we used Tableau Professional (Tableau Software, Seattle, WA) for data visualization.

RESULTS

Simulation study

Selected results for the simulations with an exposure prevalence of 30% are shown in Table 1 and full results for all exposure prevalences in Appendix Tables A.1–A.5. At a 30% exposure prevalence, we observed a mean bias of 943% (a treatment effect of 10.43 instead of 1.00) in the unmatched cohort.

Across our simulations, we observed the following trends:

- The estimates, bias, and variance when employing the within-set primary analytic method and the across-sets secondary method were virtually identical. Results

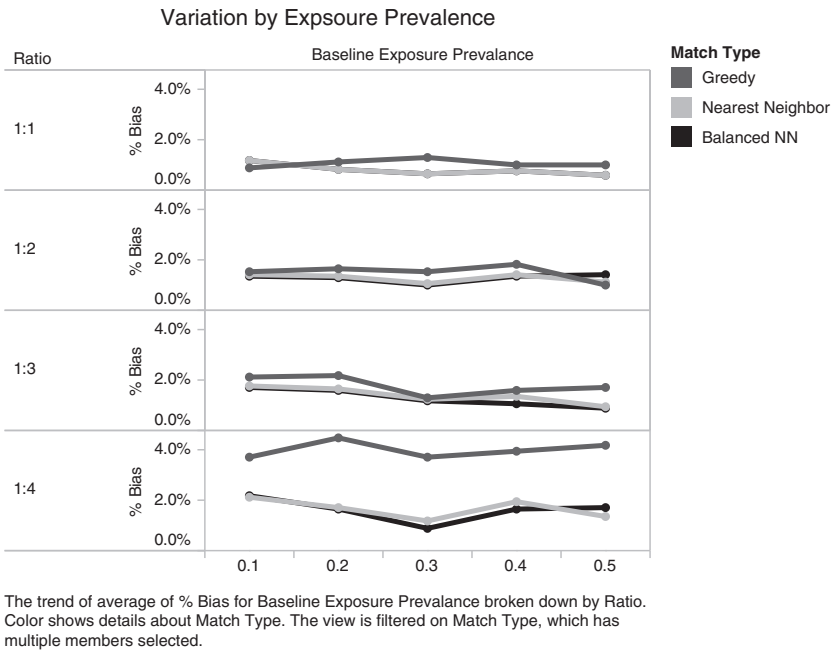

Figure 3. Average bias observed by matching type, as observed at increasing levels of expected exposure prevalence. The medium shaded line indicates greedy matching; the lighter shaded line indicates nearest neighbor; and the darker shaded line indicates balanced nearest neighbor

DOI: 10.1002/pds

from the primary method are reported in the succeeding text.

- The smallest mean covariate distance across all scenarios was observed in the 1:1 matching scenarios; this is to be expected as the best matches will occur first (Table 1). This smallest distance did not equate to the lowest bias ($r = 0.24$ for the correlation between distance and bias), because in our design, some variables had a stronger association with outcome than others.
- Increasing the match ratio beyond 1:1 generally resulted in somewhat higher bias (Figure 1).
- Increasing the match ratio beyond 1:1 resulted in lower variance using variable ratio matching but often sharply higher variance with fixed ratio matching. We attribute the difference between variable and fixed to variable ratio matching's substantially larger cohorts, as requiring fixed ratio matches results in a loss of matched sets (Figure 1).
- Cohort sizes were substantially similar among the matching methods (Figure 1).
- With variable ratio matches, MSE was substantially similar among the three matching methods. The sequential approach had a slightly lower MSE than the parallel approach, even though the bias was lower in the parallel approach. With 1:1 matches, greedy matching had the lowest MSE but also the highest bias (Figure 1).
- The parallel approach combined with variable ratio matching always yielded smaller covariate distances and biases than the sequential approach but also somewhat smaller cohorts (Figure 1).
- Overall, the largest biases were observed in the greedy matching scheme with the higher matching ratios (1:3 and 1:4). The lowest biases were observed in the balanced nearest neighbor scheme (Figures 1 and 2).
- Bias did not vary significantly with exposure prevalence (Figure 3).
- Compared with 1:1 matching, balanced nearest neighbor matching applied in parallel with a variable matching ratio had similar bias and lower MSE (Figure 1).
- With our re-implemented version of greedy matching, computing time among the three types of matching was substantially similar, with match times of 1–2 seconds for 20 000 patients.

Empirical studies

There were substantial differences between initiators of coxibs ($n = 32\,042$) and initiators of non-selective NSAIDs ($n = 17\,611$); coxib initiators were on average older (mean age 80 vs 78 years), more likely to be women (85.9% vs 81.2%), and more likely to have a history of

Table 2. Selected results of matching in a cohort of initiators of a cyclo-oxygenase 2 inhibitor and initiators of non-selective NSAIDs, with an outcome of upper gastrointestinal bleed within 120 days

Matching scheme	Standardized differences of selected measured variables						
	Age	Female	Charlson Score	Hypertension	Gastroprotective Drug	Warfarin	Number of Medications
Unmatched	0.268	0.130	0.100	0.058	0.161	0.216	0.189
1:1 Matching							
Nearest neighbor matching	0.003	−0.009	0.012	–0.003	0.011	0.013	0.014
Digit-based greedy matching	0.002	−0.004	0.011	−0.006	0.015	0.015	0.018
1:3 Matching							
Nearest neighbor matching							
Sequential variable ratio	0.051	0.022	0.020	0.002	0.031	0.030	0.041
Parallel variable ratio	0.090	0.033	0.042	0.007	0.056	0.067	0.060
Sequential fixed ratio	−0.090	0.074	0.130	−0.078	0.176	0.199	−0.118
Parallel fixed ratio	0.024	−0.061	0.067	−0.007	0.034	0.001	0.051
Balanced nearest neighbor matching							
Sequential variable ratio	0.051	0.022	0.020	0.002	0.031	0.030	0.041
Parallel variable ratio	0.079	0.031	0.035	0.006	0.047	0.056	0.056
Sequential fixed ratio	−0.059	0.062	0.117	−0.104	0.080	0.191	−0.130
Parallel fixed ratio	−0.042	−0.034	0.065	0.034	0.026	−0.005	0.083
Digit-based greedy matching							
Sequential variable ratio	0.048	0.025	0.018	0.000	0.032	0.029	0.040
Parallel variable ratio	0.068	0.022	0.034	0.004	0.049	0.059	0.048
Sequential fixed ratio	−0.066	−0.045	0.102	−0.048	0.177	†	−0.029
Parallel fixed ratio	0.062	−0.058	0.102	0.009	−0.004	0.017	0.074

NA, not applicable.
*The maximum observed standardized difference among all measured variables.
†The difference was not computable or the model did not converge.

DOI: 10.1002/pds

ulcer (3.7% vs 2.4%; Appendix Table A.6). The observed mean standardized distance between treatment groups before propensity score matching was 0.091. As compared with other statin initiators ($n = 244\,916$), initiators of a statin plus ezetimibe ($n = 13\,280$) were younger (mean age 73 vs 75 years), less likely to be women (32.9% vs 37.0%), and had fewer cardiovascular-related diagnoses (mean 3.5 vs 4.1 diagnoses; Appendix Table A.7). The observed mean standardized distance between treatment groups before propensity score matching was −0.091.

1:*n* matching was challenging in the coxib study as there were many more coxib (treated) patients than non-selective NSAID (comparison) patients. Indeed, with 1:2 fixed ratio matching, only 3% of treated patients were successfully matched to two untreated patients (Table 2). Parallel, variable ratio matching allowed for a mean matching ratio of up to 1:1.5, but the parallel approach excluded certain patients who appeared in the original 1:1 match, owing to the "starvation" phenomenon mentioned previously; the mean percentage of treated patients matched fell from 51% to 36%. Comparing 1:3 parallel, variable ratio matching with 1:1 matching, both using nearest neighbor, the observed odds ratio was 0.93 (95% confidence interval (CI) 0.73, 1.18) versus 0.88 (95% CI 0.70, 1.09), reflecting the changing population of matched patients. One solution to the problem of the matching ratio would have been to reverse the treated and untreated groups and thus achieve a 2:1 match of two coxib patients to a single non-selective NSAID patient.

1:*n* matching was more successful in the statin study as there were approximately 14 comparison patients available for each treated patient. In almost all cases, 100% of treated patients were matched to up to five comparison patients. With 1:1 nearest neighbor matching, the observed mean of the standardized differences was 0.006 (Table 3); this varied from 0.000 to 0.009 in the various scenarios considered. In the 1:1 nearest neighbor case, the observed odds ratio was 0.82 (95% CI 0.73, 0.93). With 1:5 variable ratio, parallel, balanced, nearest neighbor matching—an approach that appeared favorable in the simulation studies—the average matching ratio achieved was 1:4.5, and the observed odds ratio was 0.86 (95% CI 0.78, 0.96). Whereas the point estimate shifted upward, the 95% confidence interval also narrowed.

DISCUSSION

We investigated various approaches to 1:*n* matching in cohort studies, including a commonly used greedy matching technique, pairwise nearest neighbor matching with a caliper, and a balanced pairwise nearest neighbor approach. In our simulation

Number of doctor visits	Mean standardized distance	Max. standardized distance*	Mean number of matched sets	Mean % of treated patients matched	Mean matching ratio	Observed risk difference × 100 (95%,CI)	Observed odds ratio (95%,CI)
0.130	0.091	0.304	32 042	NA	NA	0.09 (–0.10, 0.29)	1.09 (0.60, 1.97)
0.004	0.003	0.014	16 423	51	1.0	−0.13 (−0.29, 0.02)	0.88(0.70, 1.09)
0.007	0.004	0.018	16 561	52	1.0	−0.14 (−0.30, 0.01)	0.87 (0.70, 1.07)
0.026	0.013	0.097	16 423	51	1.1	−0.11 (−0.22, −0.00)	0.87 (0.70, 1.08)
0.045	0.029	0.090	11 619	36%	1.5	−0.04.(−0.16, 0.08)	0.93 (0.73, 1.18)
0.260	0.055	0.260	101	0	3.0	−0.33 (−0.89, 0.23)	†
0.080	0.022	0.080	1736	5	3.0	−0.15 (−0.41, 0.11)	0.84.(0.46, 1.52)
0.025	0.013	0.096	16 423	51	1.1	−0.11 (−0.22, −0.00)	0.87.(0.70,.1.08)
0.039	0.024	0.080	13 464	42	1.3	−0.03 (−0.14, 0.08)	0.95.(0.76,.1.20)
0.266	0.048	0.266	110	0	3.0	−0.61 (−1.33, 0.12)	†
0.121	0.032	0.121	1411	4	3.0	0.05 (−0.25, 0.34)	1.05 (0.57, 1.95)
0.025	0.012	0.098	16 561	52	1.1	−0.11 (−0.22, −0.01)	0.86 (0.70, 1.07)
0.033	0.022	0.068	12 564	39	1.4	−0.10 (−0.22, 0.01)	0.89 (0.70, 1.12)
0.364	0.061	0.364	75	0	3.0	−0.44 (−1.20, 0.31)	†
0.132	0.024	0.132	1312	4	3.0	0.03 (−0.27, 0.32)	1.03 (0.53, 1.99)

DOI: 10.1002/pds

Table 3. Selected results of matching in a cohort of concurrent initiators of a statin plus ezetimibe or combination product (Vytorin) and initiators of any other statin, with an outcome of myocardial infarction, stroke, or death within 180 days

Matching scheme	Standardized differences of selected measured variables														
	Age	Female	Charlson Score	Hypertension	Diabetes	Number of medications	Hosiptalizations with CVD diagnoses	CV drugs	Mean std. distance	Max. std. distance*	Mean number of matched sets	Mean % of treated patients matched	Mean matching ratio	Observed risk difference × 100 (95%,CI)	Observed odds ratio(95%,CI)
Unmatched	−0.180	0.037	−0.019	0.116	0.061	0.100	−0.092	0.046	0.018	0.454	22793	NA	NA	−11.29 (−1.52,.−1.05)	0.60 (0.39, 0.93)
1:1 Matching															
Nearest neighbor matching	0.035	0.006	0.019	0.004	0.010	0.006	0.015	0.006	0.006	0.035	22 793	100	1.0	−0.42 (10.61,.10.23)	0.82 (0.73, 0.93)
Digit-based greedy matching	0.006	0.005	0.014	−0.006	0.005	0.004	0.010	−0.006	0.002	0.028	22 793	100	1.0	−0.24 (−0.43, −0.05)	0.89 (0.78, 1.01)
1:5 Matching															
Nearest neighbor matching															
Sequential variable ratio	0.015	−0.001	0.009	0.000	0.005	0.011	0.009	0.005	0.004	0.021	22 793	100	5.0	−0.27 (−0.36, −0.19)	0.88 (0.79, 0.97)
Parallel variable ratio	0.016	−0.001	0.009	0.000	0.005	0.010	0.010	0.005	0.004	0.022	22 779	100	5.0	−0.27 (−0.36, −0.18)	0.88 (0.80, 0.97)
Sequential fixed ratio	0.018	−0.001	0.008	0.000	0.005	0.010	0.010	0.004	0.003	0.022	22 572	99	5.0	−0.27 (−0.36,−0.18)	0.88 (0.80, 0.97)
Parallel fixed. ratio	0.017	−0.002	0.008	0.000	0.005	0.008	0.010	0.004	0.003	0.022	22 714	100	5.0	−0.27 (−0.36, −0.18)	0.88 (0.79, 0.97)
Balanced nearest neighbor matching															
Sequential variable ratio	0.010	−0.001	0.006	−0.001	0.004	0.008	0.006	0.004	0.003	0.030	22 793	100	5.0	−0.30 (−0.39, −0.21)	0.87 (0.79, 0.96)
Parallel variable ratio	−0.011	0.003	0.001	0.005	0.005	0.011	−0.004	0.007	0.003	0.080	22 793	100	4.5	−0.37 (−0.46, −0.29)	0.86 (0.78, 0.96)
Sequential fixed ratio	0.013	−0.002	0.004	−0.002	0.001	0.003	0.008	0.003	0.002	0.025	22 165	97	5.0	−0.29 (−0.38, −0.20)	0.87 (0.79, 0.97)
Parallel fixed ratio	0.003	−0.003	−0.004	−0.011	−0.002	0.000	0.003	0.007	0.000	0.029	18 135	80	5.0	−0.32 (−0.42, −0.22)	0.87 (0.78, 0.97)
Digit-based greedy matching															
Sequential variable ratio	0.012	0.003	0.007	−0.004	0.002	0.006	0.004	0.004	0.002	0.023	22 793	100	5.0	−0.32 (−0.41, −0.23)	0.86 (0.78, 0.95)
Parallel variable ratio	−0.013	0.010	0.012	0.018	0.013	0.026	−0.001	0.008	0.009	0.128	22 793	100	4.5	−0.36 (−0.45, −0.27)	0.88 (0.79, 0.97)
Sequential fixed ratio	0.012	−0.001	0.005	−0.005	−0.003	0.004	0.004	−0.001	0.001	0.026	22 734	100	5.0	−0.27 (−0.36, −0.19)	0.88 (0.79, 0.97)
Parallel fixed ratio	0.004	−0.009	0.003	−0.003	0.002	−0.003	0.001	0.006	0.000	0.024	18 237	80	5.0	−0.32 (−0.42, −0.22)	0.87 (0.78, 0.97)

*The maximum observed standardized difference among all measured variables.

DOI: 10.1002/pds

analysis, we observed that variable ratio matching consistently outperformed fixed ratio matching with respect to bias, precision, and MSE and that 1:*n* variable ratio matching yielded higher precision than 1:1 matching at the cost of a small increase in bias. Of the variable ratio approaches, our pairwise nearest neighbor and balanced nearest neighbor approaches both resulted in lower bias than the commonly used greedy matching approach. Whereas the precision of 1:1 matching may suffice in many cases, particularly if transparency with respect to balance achieved is of importance, we observed that variable ratio, parallel, balanced, 1:*n* nearest neighbor matching can be used to increase precision at little cost in bias. We provide the software in our Pharmacoepidemiology Toolbox (http://www.hdpharmacoepi.org).

The transparency issue with variable ratio matching can be substantial. Owing to the differing numbers of patients in each matched set, a simple "Table 1" of a variable ratio matched cohort will not show how well covariate balance was achieved. We see three approaches, none ideal, to presenting a Table 1 in this situation. First, one could present a single Table 1 with each matched set's single best match. This has the advantage of displaying balance but the disadvantages of (i) not showing the entire cohort that contributed to the analysis and (ii) showing only the maximal balance, because additional matches will generally be of lower quality. Second, one could present a series of Table 1, one for each matched set size (a Table 1 of the 1:1 matches, the 1:2 matches, and so forth). Although this will display the entire cohort, it may confuse readers and could be perceived as excessive. Finally, each patient's contribution to a single Table 1 could be weighted by matched set size; this approach will illustrate balance but is not a pure display of the data. A hybrid approach, with both the 1:1 "best matches" displayed alongside a weighted population, could also be feasible.

This study considered only continuous outcomes; a limited simulation study that we performed indicated that the results should be substantially similar with dichotomous outcomes, but further investigation is required. Further investigation should also consider the use of *n:m* matching, a form of fine stratification, as described in prior literature.[14]

Whereas 1:1 matching may yield sufficiently precise estimates in large studies or studies with strong effects, we find that variable ratio, parallel balanced, 1:*n* nearest neighbor matching was a reasonable way to improve precision with little cost in bias but did come with a loss of transparency. Depending on the sizes of exposed and comparison populations and the need for precision, picking an appropriate matching strategy can optimize results.

CONFLICT OF INTEREST

Dr. Rassen is a recipient of a career development award from Agency for Healthcare Research and Quality (K01 HS018088). The Division of Pharmacoepidemiology received gifts from IBM Netezza and Tableau Software. Dr. Glynn served as consultant to Merck, and in that capacity he gave a methods-oriented talk at the company. The topics of the talk included Vioxx but were unrelated to this study.

Supporting Information

Additional supporting information may be found in the online version of this article:

Appendix Table 1: Matching simulation results for base exposure prevalance of 10%

Appendix Table 2: Matching simulation results for base exposure prevalance of 20%

Appendix Table 3: Matching simulation results for base exposure prevalance of 40%

Appendix Table 4: Matching simulation results for base exposure prevalence of 50%

Appendix Table 5: Baseline characteristics of a cohort of initiators of cyclo-oxygenase 2 inhibitors and initiators of non-selective NSAIDs

Appendix Table 6: Baseline characteristics of a cohort of concurrent initiators of a statin plus ezetimibe or combination product Vytorin) and initiators of any other statin

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

REFERENCES

1. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd edn. Lippincott Williams & Wilkins: Philadelphia, 2008.
2. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* 1997; **127**: 757–63.
3. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
4. D'Agostino RB, Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* 1998; **17**: 2265–81.
5. Miettinen O. Stratification by a multivariate confounder score. *American Journal of Epidemiology* 1976; **104**: 609–20.

6. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer Stat* 1985; **39**: 33–8.
7. Imbens G. Nonparametric estimation of average treatment effects under exogeneity: a review. *The Review of Economics and Statistics* 2004; **86**: 4–29.
8. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics* 2011; **10**: 150–161.
9. Sturmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. *American Journal of Epidemiology* 2010; **172**: 843–54.
10. Reducing bias in a propensity score matched-pair sample using greedy matching techniques. 2001. (Accessed at www2.sas.com/proceedings/sugi26/p214-26.pdf.)
11. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine* 2007; **26**: 3078–94.
12. Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. *Biometrical Journal* 2009; **51**: 171–84.
13. Austin PC. Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *American Journal of Epidemiology* 2010; **172**: 1092–7.
14. Ming K, Rosenbaum PR. Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics* 2000; **56**: 118–24.
15. Performing a 1:N case–control match on propensity score. 2004. (Accessed at http://www2.sas.com/proceedings/sugi29/165-29.pdf.)
16. SAS Macros. (Accessed December 15, 2011, at http://mayoresearch.mayo.edu/mayo/research/biostat/sasmacros.cfm.)
17. Sturmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol* 2006; **59**: 437–47.
18. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine* 2008; **27**: 2037–49.
19. MatchIt: nonparametric preprocessing for parametric causal inference. 2011. (Accessed at http://gking.harvard.edu/matchit.)
20. Hill J. Discussion of research using propensity-score matching: comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. *Statistics in Medicine* 2008; **27**: 2055–61; discussion 66–9.
21. Cormen TH. *Introduction to Algorithms* 3rd edn. MIT Press: Cambridge, Mass, 2009.
22. Brookhart MA, Rassen J, Wang PS, Dormuth CA, Mogun H, Schneeweiss S. Evaluating the validity of an instrumental variable study of neuroleptics: can between-physician differences in prescribing patterns be used to estimate treatment effects? *Medical Care* 2007; **45**: S116–S22.
23. Schneeweiss S, Solomon DH, Wang PS, Rassen J, Brookhart MA. Simultaneous assessment of short-term gastrointestinal benefits and cardiovascular risks of selective cyclooxygenase 2 inhibitors and nonselective nonsteroidal antiinflammatory drugs: an instrumental variable analysis. *Arthritis and Rheumatism* 2006; **54**: 3390–8.
24. Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* 2006; **17**: 268–75.
25. Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables II: in 25 variations, the physician prescribing preference generally was strong and reduced imbalance. *Journal of Clinical Epidemiology* 2009; **62**: 1233–41.
26. Rassen JA, Doherty M, Huang W, Schneeweiss S. Pharmacoepidemiology Toolbox version 2. In. Boston, MA; 2011.

APPENDIX A

IMPROVEMENTS IN THE EFFICIENCY OF NEAREST NEIGHBOR MATCHING ALGORITHMS

To accommodate our simulations, we re-implemented nearest neighbor matching to execute in $O(n \log n)$ time; that is, the algorithm will require order of magnitude n times the log of n iterations to execute. This is far faster than the usual $O(n^2)$ "brute force" algorithm. In principle, we achieve this by taking advantage of the fact that a treated patient's nearest referent match can only be the referent patient with the next highest ("immediately to the right") or the next lowest ("immediately to the left") propensity score. The algorithm executes as follows:

1. With each treated patient, construct two best possible matches: the comparison patient immediately to the left and the comparison patient immediately to the right. Calculate the distance for each match, and place both on a list of possible matches sorted by lowest to highest distance (the "heap").
2. Take the top item off the heap. This will be the single best match (lowest distance) in the population. Place this match in the final match list. Mark both patients (the treated patient and the comparison patient) as matched.
3. Take the new top item off the heap. If neither patient in the potential match has already been matched, place the match in the final match list and mark the patients as matched. If both patients have been matched, then ignore the pair; these patients have already been matched in more optimal pairings. If one of the two patients has been matched, then that patient has to be replaced by his next-best alternative; the best alternative will be the next "unmatched" patient immediately to the right or left. Make the replacement, calculate the match distance, and place this new matched set back in the sorted heap.
4. Repeat Step (3) until no more matches are in the heap.

APPENDIX B

PAIRWISE NEAREST NEIGHBOR VERSUS OPTIMAL MATCHING

It is possible that nearest neighbor matches will not yield globally optimal matches; consider the figure. With pairwise nearest neighbor matching, the first matched pair in the cohort will be the pair of patients with the smallest inter-patient distance, the second matched pair will be the patients with the smallest distance among the remaining patients, and so forth. The pairwise nearest neighbor match would yield pairings of ($r2$, $b1$) and ($r1$, $b2$), even though a global optimization would select (r_1, b_1) and (r_2, b_2) and thus yield a lower total match distance. The pairwise nearest neighbor scheme trades off average match quality for the best possible single pairings, up to a certain maximum allowable match distance (the caliper). Whereas this theoretically appears to be a trade-off, practically there appears to be little difference.

DOI: 10.1002/pds