PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2012; 21(S1): 41–49
Published online in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/pds.2328

ORIGINAL REPORT

Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system

Jeremy A. Rassen* and Sebastian Schneeweiss

Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

ABSTRACT

Distributed medical product safety monitoring systems such as the Sentinel System, to be developed as a part of Food and Drug Administration's Sentinel Initiative, will require automation of large parts of the safety evaluation process to achieve the necessary speed and scale at reasonable cost without sacrificing validity. Although certain functions will require investigator intervention, confounding control is one area that can largely be automated. The high-dimensional propensity score (hd-PS) algorithm is one option for automated confounding control in longitudinal healthcare databases. In this article, we discuss the use of hd-PS for automating confounding control in sequential database cohort studies, as applied to safety monitoring systems. In particular, we discuss the robustness of the covariate selection process, the potential for over- or under-selection of variables including the possibilities of M-bias and Z-bias, the computation requirements, the practical considerations in a federated database network, and the cases where automated confounding adjustment may not function optimally. We also outline recent improvements to the algorithm and show how the algorithm has performed in several published studies. We conclude that despite certain limitations, hd-PS offers substantial advantages over non-automated alternatives in active product safety monitoring systems.

KEY WORDS—propensity scores; confounding factors (epidemiology); multicenter study (publication type); epidemiological methods

INTRODUCTION

Distributed safety monitoring systems, such as the Sentinel System to be developed as a part of Food and Drug Administration's Sentinel Initiative, will benefit from automating large parts of today's pharmacoepidemiology study process and will require automation that is based on sound design principles and careful quality control, and built to ensure validity while providing speed and scale. Speed is required for fast evaluation and identification of safety signals; scale is needed to manage both the number of patients under observation and the number of potential safety signals the system needs to be able to evaluate. In this setting, the intelligence that investigators normally apply study-by-study needs to be encapsulated in study frameworks and algorithms that can be applied reliably and in a largely hands-off fashion.

The choice of an appropriate study design can largely be driven by attributes of the safety question and associated appropriate designs. Self-controlled designs, which minimize confounding by comparing a patient with himself or herself, are well suited to studies of acute onset events such as allergic reactions or situations of transient exposures.[1] However, in most monitoring scenarios, cohort designs and their associated sampling strategies will be better suited. A successful safety cohort study will use an incident user design[2] with well-defined covariate assessment and exposure definition windows, covariate balancing, and other familiar components.[3] An important part of cohort-type designs is choosing a suitable comparison group. This choice is key to study validity and will need substantial investigator input on a case-by-case basis.

In this article, we have made the following assumptions: (i) we will use a cohort-type design with the understanding self-controlled designs are not applicable to the monitoring setting at hand;[4] (ii) for practical reasons, data updates occur at predefined, periodic time points (e.g. every 3 months) and the cohort is

*Correspondence to: J. A. Rassen, Department of Medicine, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. E-mail: jrassen@post.harvard.edu sequentially expanded in size; (iii) a drug with a similar indication as the study drug has been chosen as a clinically-meaningful active comparator; and (iv) the study data are longitudinal insurance claims stored across a distributed database network.

Cohort studies of treatment effects require control for between-person confounding. Common methods to minimize confounding bias in non-randomized cohort include inducing homogeneity in patient characteristics by restriction of the patient population, stratification of treatment effects by subgroups, matching patients on key factors such as age, or adjusting for measured confounders with regression. Propensity scores and disease risk scores extend these basic concepts by aggregating a number of confounders into a single summary measure.

Any control of confounding begins with the identification of potential confounding factors and the correct selection of the covariates that influence the use of study medications and the outcome under evaluation. Traditionally, this is done through the application of subject matter expertise or through a more formal process such as directed acyclic graphs.[5] Covariates may be created (identified and measured) specifically for the study, or in pharmacoepidemiological database studies, covariate definitions may already exist in a standing library. However, this traditional approach does not scale well to either the large number of covariates necessary for reasonable confounding control in pharmacoepidemiological database studies[6] nor to the number of monitoring projects envisioned for an active medical product surveillance system. A standing library of covariate definitions, even one based on a common data model, may not include all the important risk factors for the present study and may require all participating data partners to subscribe to a 'lowest common denominator' of available data elements. The fact that drug user populations shift over time will make it even more cumbersome to pick a single set of 'correct' variables that are applicable over time.

The ideal automated procedure for covariate selection would create and select pre-exposure covariates and, by controlling for these covariates, minimize residual bias as well as or better than a team of investigators would be able to. This process would enable the valid study of medical product-outcome associations with a minimum of investigator intervention and thus make monitoring fast and scalable.

In this article, we investigate whether one procedure, the high-dimensional propensity score (hd-PS) algorithm, could serve as an automated mechanism for confounding adjustment for the Sentinel System and describe the strengths and limitations of hd-PS in such a setting. In particular, we discuss the robustness of the covariate selection process, the computation required, the potential for over- or under-selection of variables, the practical considerations in a federated database network, and the cases where automated confounding adjustment may not function optimally.

THE HD-PS ALGORITHM

In earlier work,[6,7] we described the hd-PS algorithm, an automated covariate creation, selection, and confounding adjustment process. It has now been applied to several pharmacoepidemiology studies.[8–13] Moving left to right across each row of Table 1 shows what we consider to be evidence of the method's success: across a range of studies, we see a largely monotonic progression of the point estimate as additional levels of confounding control are applied. We observed that this progression may move toward a null finding, away from the null, or even to and beyond the null depending on the nature of the residual confounding. For example, in the first row, the unadjusted point estimate indicate that Cox-2 inhibitors (coxibs) are associated with a 9% increase of incidence of gastrointestinal bleed as compared with non-selective non-steroidal anti-inflammatory drugs (ns-NSAIDs). Randomized trials suggested a lowering of risk by approximately 20% among healthy patients,[14,15] so we consider the unadjusted value, observed among patients undergoing routine care, to be upwardly biased. Adjusting for age, sex, and other basic variables moves the point estimate downward to a 1% increase in risk. Further adjustment by other investigator-specified variables moves the estimate further downward, to a 6% relative risk reduction. Applying hd-PS moves the estimate yet further downward, to a 12% to 13% relative risk reduction observed in our routine care population.

Since the algorithm's original publication, we have studied and made modifications to the procedure both to handle small study sizes[6] and to greatly improve speed.[16]

APPLYING THE HD-PS ALGORITHM IN THE SENTINEL SYSTEM

The hd-PS algorithm creates and prioritizes potential confounders of the medical product-outcome association under study.[7] In its most common configuration, it takes as input the recorded history of medical encounters—the presence of diagnostic codes, procedure codes, hospitalizations, and medication fills—experienced by the patient before exposure. The algorithm creates covariates from each of these events—coded for presence or absence of the event

Table 1. hd-PS as applied to various studies

			Relative risk estimate (95%CI)				
Type of data source	Exposure (referent)	Outcome	Unadjusted	Adjusted by basic variables	Basic variables plus other pre-specified variables	Basic variables plus hd-PS	Basic variables and hd-PS Only
US Medicare claims data	Coxibs (ns-NSAIDS) [7]	GI bleed within 180 days	1.09 (0.91–1.30)	1.01 (0.84–1.21)	0.94 (0.78–1.12)	0.88 (0.73–1.06)	0.87 (0.72–1.05)
German claims data	Coxibs (ns-NSAIDS) [41]	GI bleed	1.21 (0.91–1.61)	–	0.99 (0.74–1.33)	0.67 (0.45–0.97)	–
UK claims data supplemented with EMR data	Coxibs (ns-NSAIDS)[13]	Upper GI bleed	1.50 (0.98–2.28)	0.84 (0.54–1.31)	0.81 (0.52, 1.27)	0.78 (0.49, 1.22)	0.81 (0.52–1.28)
US Medicare claims data	Statins (glaucoma drugs)[7]	All-cause mortality within 180 days	0.56 (0.51–0.62)	0.77 (.069–0.85)	0.80 (0.70–0.90)	0.86 (0.76–0.98)	0.89 (0.78–1.02)
British Columbia claims data	TCAs (SSRIs), <18 y.o.[9]	Suicide within 1 year	0.59 (0.28–1.27)	0.66 (0.31–1.42)	0.71 (0.33–1.52)	0.92 (0.43–2.00)	–
British Columbia claims data	TCA (SSRIs), 18+ y.o.[10]	Suicide within 1 year	0.97 (0.77–1.21)	1.04 (0.83–1.31)	1.04 (0.82–1.31)	1.14 (0.88–1.47)	–
Mix of US Medicare, US commercial, and Canadian claims data	Clopidogrel + PPI (clopidogrel alone)[42]	MI or CV death	1.74 (1.44–2.10)	1.62 (1.34–1.96)	1.32 (1.08–1.61)	1.22 (0.99–1.51)	–
US commercial claims data	Neurontin (Topamax)[11]	Suicide or attempted suicide	0.95 (0.76–1.19)	1.48 (1.17–1.87)	1.42 (1.11–1.80)	1.99 (1.45–2.73)	–
British Columbia claims data	Conventional APMs (atypical APMs)[12]	Death	1.37 (1.11–1.69)	1.47 (1.13–1.90)	1.47 (1.14–1.91)	1.52 (1.14–2.02)	–
British Columbia claims data	Benzodiazepines (atypical APMs)[12]	Death	1.37 (1.14–1.64)	1.52 (1.25–1.85)	1.28 (1.04–1.58)	1.20 (0.96–1.50)	–

One measure of performance of the hd-PS algorithm is a consistent movement in point estimates as more confounding adjustment is applied (right to left in the table). TCA, tri-cyclic anti-depressant; SSRI, selective serotonin uptake inhibitor; PPI, proton pump inhibitor; APM, anti-psychotic medication; MI, myocardial infarction; CV, cardiovascular; GI, gastrointestinal.

DOI: 10.1002/pds

and, when present, for frequency of the event's recurrence—and assesses these new covariates for their association with both study exposure and outcome. Using Bross' formula for confounding bias for dichotomous variables,[17] it then ranks the group of covariates for their potential to bias the association under study, and by default will select the top 500 of these covariates that seem most likely to add bias. It enters these variables into an exposure propensity score model, and after estimating the propensity score, the hd-PS algorithm initiates a fixed-ratio matching process that creates a cohort in which patients treated with the exposure and referent drugs are balanced with respect to measured covariates. Although matching may implicitly exclude patients who are in the tails of the propensity score distributions,[18] it will estimate the treatment effect only among patients who could plausibly have received either of the drugs under study. In randomized trial terms, these are the patients for whom there was equipoise.

When using hd-PS, we recommend a set of basic design principles. To ensure clear temporality and to avoid other biases, we apply an incident user cohort design[3] in which exposure is required to be preceded by a term of non-use of the study drugs, thereby excluding prevalent users. In most cases, exposure should be contrasted with an active comparator group with the same indication; for example, users of coxibs would be compared with users of ns-NSAIDs rather than non-users.[2] All covariates that are assessed must be recorded within a defined period before the exposure date, often 180 or 365 days. Outcome is assessed during a limited follow-up time with censoring at the time of treatment discontinuation or treatment group switching. These design criteria, while not the only way to conduct a successful study, ensure that key epidemiological principles are met: covariates are measured before exposure, incident users are compared 'apples to apples' to other incident users, and exposure misclassification during follow-up is limited.

With this in place, hd-PS can identify and select covariates, estimate a propensity score, and match patients within the cohort, all without user intervention. Applied on a periodic basis over time and at each participating data provider's facility, the result is a series of cohorts matched *within* each provider's patient base[19] and matched with respect to the best available covariate information at the time of the patients' exposures. As exposure or outcome frequency grows or as the composition of the population using a drug changes over time, so too will the covariates selected to optimally address confounding. Although this is contrary to the principle of choosing covariates based on knowledge of the biologic processes at work, it is an effective approach in secondary data like those that will be used within the Sentinel System. It is also pragmatic, as it addresses the issue of changing drug usage patterns over time and thus ensures that maximal validity is achieved at each point in time and in each data environment, even across the heterogeneous data elements available in a federated system.

As we noted earlier, matching the cohorts imposes an implicit and useful restriction criterion: patients are unmatched and thus excluded if there are exposed patients for whom no exchangeable referent patient exists, or *vice versa*. Although analytically the results are similar to a trimmed propensity score approach,[18] a fixed-ratio matching process also provides other beneficial side effects, such as a cohort that should be balanced with respect to all measured confounders and thus does not require further confounding control in the outcome analysis. An inspection of the cohort stratified by exposure category and confounders will indicate, either visually or through the automated application of measure like the Mahalanobis distance,[20,21] residual imbalances that need to be addressed. The benefits of the simplicity of this balance verification should not to be underestimated in a system that aspires to automate as many aspects as possible but which still allows for rapid quality checks of the underlying epidemiology.

DISCUSSION OF THE STRENGTHS AND LIMITATIONS OF HD-PS

For decades, epidemiologists have been taught that each confounding variable's relationship with the exposure and outcome must be fully understood on biological and medical-sociological grounds. Unsurprisingly, a healthy skepticism is a common first reaction to an automated confounding adjustment approach. However, context is important: in the case of a safety surveillance system based on secondary data, our biggest concern is unmeasured confounders, as we are not in control of the data collection process and might not know all relevant confounders and thus are prevented from defining necessary confounding variables from first principles.

In this section, we present several of hd-PS's performance characteristics as well as potential threats to validity because of the use of the hd-PS algorithm.

General issues with selection of variables for confounding adjustment

An automated variable selection technique can fail in several ways: it can select too few covariates, too many covariates, and/or can select the wrong covariates.

With respect to the optimal number of covariates to be selected, we recently performed extensive simulation studies in which we sought to determine how many variables are required for maximal confounding adjustment achievable in a specific data source for a specific exposure-outcome pair.[6] We examined hd-PS in both common and small study circumstances and determined that 350 to 400 variables are generally sufficient. A additional variables provided a little change in point estimate, as the algorithm first selects the variables likely to cause the most bias. Selecting additional variables would likely do no harm, but selecting fewer is likely to lead to under-adjustment. Indeed, in a propensity score, there is little harm to including too many variables, as long as those variables are either confounders, proxies for confounders, or are predictors of the outcome.[22] Further, because the parameters of a propensity score model will not be interpreted, and the score will not be generalized to other data sets, there is no need for concern about overfitting.[23,24] Naturally, there will be a strong correlation among certain measures (e.g. ordering of a cholesterol test, diagnosis of hypercholesterolemia, statin drug use, and increased service usage intensity), but while this correlation would affect the interpretation of the elements of the propensity score model, it will not reduce the score's ability to adjust for confounding.

Selection of the *correct* variables is nevertheless important. Although the Bross formula provides a reasonable way to quantify the potential univariate bias if we would fail to adjust for a particular variable, it does not account for more complex situations. In particular, certain variables such as colliders[25,26] may seem analytically to be confounders, but adjusting for them can increase bias.[5]

Two relevant kinds of collider bias have been noted in the literature: M-bias, named for the shape of the directed acyclic graph that characterizes it,[26] and Z-bias, named because the bias comes from the inclusion of an instrumental variable (often notated Z) in the analysis; Z-bias is, also called 'residual confounding amplification'.[27,28]

Variable selection issue: potential for M- and Z-bias

M-bias (Figure 1) occurs from conditioning on an apparent confounder (C), which is actually a collider. C must be associated with two types of unmeasured confounders—a U1 that is associated only with exposure and a U2 that is associated only with the outcome—but also not be directly associated with either exposure or outcome. Brookhart *et al.*[28] gave this hypothetical example: in a study of anti-depressants' effect on incidence of lung cancer, assume that U1 is depression status (affecting anti-depressant use but not lung cancer) and U2 is smoking history (affecting lung cancer but not anti-depressant use). By conditioning on cardiovascular disease (C), an association is induced between anti-depressants and lung cancer via the M-shaped pathway via depression, cardiovascular disease, and smoking.[5]

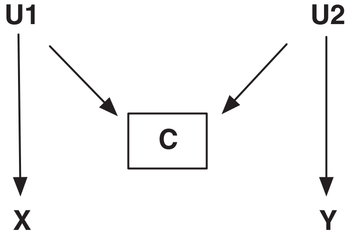

Figure 1. Example of M-bias

Although M-bias has been shown theoretically, in practice, its effect seems to be minimal. Liu *et al.*[29] showed that the bias, even when detected, was generally small (<5%, based on strength of residual confounding) and also pointed out that finding a scenario like this was rather difficult, especially in an incident user design. Moreover, outside the tidiness of a simulation environment, in other circumstances (Figure 2), the variable C may be a collider on one path (U1 → C → U2) but a proxy for a confounder (U3) on another. In this case, whether to adjust for C comes down to whether doing so will do more good than harm. It seems that in most pharmacoepidemiology situations, we are unlikely to see true M-bias of any relative magnitude,[26] and if we do, experience tells us that the complexity of the underlying biological and medical-sociological structures will likely

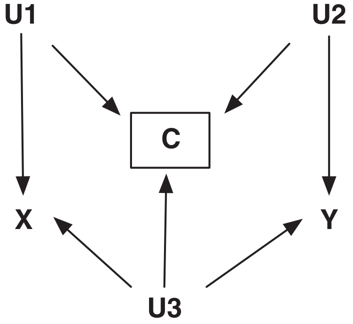

Figure 2. Example of M-bias plus confounding bias

yield a complicated situation that mixes confounding and colliding. In most of these circumstances, the reduction in bias from adjusting for the confounding should far offset any increase in bias because of conditioning on a collider.[26]

Z-bias refers to the bias caused by adjusting for an instrumental variable in studies that also have unmeasured confounding.[27,28,30,31] An instrument is a variable that is associated only with the exposure and not with outcome (other than through a pathway via exposure) and can serve to adjust for unmeasured confounding if handled with the proper analytic tools.[32–34] In a study of statins versus glaucoma drugs on the incidence of myocardial infarction, a variable like prior glaucoma diagnosis will be strongly predictive of whether a patient receives a glaucoma drug but will have little effect on the outcome.

Myers *et al*.[30] undertook a simulation study to quantify the effects of Z-bias in common pharmacoepidemiology settings. They found that Z-bias, although measureable, was only of substantial magnitude in cases of very strong unmeasured confounding, and even in these cases, the strongest Z-bias we observed represented less than 5% of the total study bias. From this simulation analysis, they concluded that Z-bias was indeed a measureable phenomenon but was small in degree compared with studies' true threat to validity: unmeasured confounding. They further conclude that when in doubt about whether a covariate is a confounder or an instrument, adjusting for the covariate will generally reduce net bias.

The first line of defense against M- or Z-bias is to reduce unmeasured confounding; by doing so, the effect of these biases will be minimized or eliminated. If unmeasured confounding remains after applying hd-PS or another confounding reduction technique, some effect of M- and Z-bias may be unavoidable: in the end, it is impossible to distinguish a confounder from a collider through inspection of data. However, in non-randomized pharmacoepidemiology, confounding bias is generally considered to be the greatest threat to study validity; any confounding bias will likely be of greater magnitude than collider bias. An automated confounding adjustment system that selects a large number of covariates, even with somewhat imperfect variable selection, should improve study validity far more than it will harm it (Figure 3).

Variable selection issue: selection with respect to outcome

Rubin[42,45] advocates that variables in a propensity score should be selected on the basis of whether the variables balance the patients between exposure groups, but not on whether they are independent risk factors for outcome. Although we agree with this approach in principle, it does assume that investigators know and measure all [*true* confounders *a priori*]. A principal innovation of hd-PS is to automatically identify a large number of covariates and prioritize them according to their potential to be confounders. Although this prioritization requires reference to the study outcome, we view hd-PS as a pragmatic approach to Rubin's general principle of using propensity scores to ultimately improve study validity.

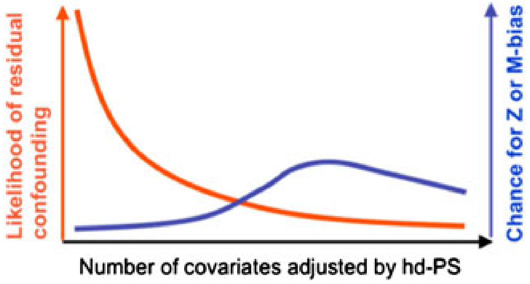

Figure 3. In most realistic scenarios, with increasing covariate adjustment, net bias should be reduced even in the theoretical presence of M- or Z-bias

Use of hd-PS with few exposed patients or outcome events

In a distributed Sentinel-type system, issues stemming from small cohort size can arise in multiple ways: it is possible that some contributing sites will be small, or that sites, although large, will have low exposure frequency. Infrequent exposure may be common in the setting of active surveillance of drugs that are new to market, as there may be a limited number of early users.[35]

With few exposed patients, propensity scores are difficult to estimate. Alternatives exist—high-dimensional disease risk scores[36] can be used in place—but in an automated system, there may be little ability to implement alternative approaches. In cases of few exposed patients and even fewer outcome events, several pragmatic approaches can be contemplated: one can wait until sufficient exposures accumulate, one can pool sites to estimate propensity scores (if data sharing is allowed), one can estimate a propensity score with a minimal number of variables and thus lower confidence in the resulting point estimate accordingly, or one can estimate a disease risk score from historical data and use that until enough exposed patients are observed.

With respect to few outcome events, we have shown that hd-PS works well with approximately 150 or more outcome events.[6] With between 25 and 150 outcome events, we recommend enabling the new 'zero cell correction' or 'confounder-exposure assessment only' options provided in the latest versions of the hd-PS algorithm.[6,36] Zero cell correction adds 0.1 to each cell of each variable-outcome 2 × 2 table and thereby enables estimation of all variables' potential bias. Alternatively, the confounder-exposure assessment mode implements the outcome-independent Rubin approach and judges potential for bias only via the variables' differential prevalence in the exposed and unexposed groups. With fewer than 25 events, we have observed that hd-PS generally works as well as investigator's specification of covariates, but may not offer any improvement beyond that.[6]

Automated generation of health utilization variables

Health service intensity variables such as number of office visits or number of medications used are important proxies for health state:[37] sicker patients have more healthcare encounters and more healthcare encounters lead to more opportunities that health status will be recorded. Because these variables are so frequently crucial confounders, the hd-PS algorithm includes automated characterization of each patient's service usage.[6,36] Early tests have shown that hd-PS's generation of health service intensity variables is equivalent to the investigator-specification of these variables, to within 1% of the resulting point estimate.

Additional challenges in a distributed database setting

The distributed data setting of the Sentinel System allows for the contribution of many participating sites with varying levels of available information, but the robust confounding control needed in most safety studies introduces certain logistical and analytic challenges. For reasons of patient and organizational privacy, a central site may not be able to receive individual-level data. A better solution may be for each site to estimate an hd-PS and then share just de-identified information—anonymous identifier, exposure status, outcome status, estimated hd-PS, and possibly other non-identifying information to be used for subgroup analyses.[38–40] The cost of this approach is a set of limitations: each participating site must have the analytic capability to run preprogrammed code for hd-PS and relevant diagnostics, including the generation of a table of the patients' characteristics as stratified by treatment group, and any subgroup analyses must be specified *a priori*. In studies to date, including a large-scale, multisite investigation that made extensive use of propensity scores,[25] the benefit of maximal confounding adjustment and thus maximal study validity has outweighed the limitations imposed.

In a distributed setting with heterogeneous data elements, hd-PS also offers two further advantages. First, the algorithm is largely agnostic to data structure and coding schemes: hd-PS has worked without modification or transformation to a common data model on data from Medicare, commercial US insurers, British Columbia's provincial insurance programs, and the UK's THIN and GPRD research databases. Second, hd-PS is designed to take maximal advantage of data available from each site; if one site has detailed, medical record-based information but another has just basic claims, hd-PS will adjust maximally in each site rather than revert to a lowest common denominator of available information. If sites have substantially differing point estimates, we have proposed methods to determine whether the variability comes from heterogeneous patient populations or from insufficient confounding adjustment in sites with less information stored in their databases.[39]

Requirements for computing time

The current version of hd-PS takes advantage of the ability of SAS 9.2 to transparently call Java programs from within data steps, and is released as a SAS macro with an embedded Java component. The hybrid SAS/Java approach allows for substantially improved performance versus a macro implemented purely in the SAS macro language. This performance is key for subgroup analyses, sensitivity analyses, or other situations in which hd-PS must be run multiple times. We have also implemented a version of hd-PS for high-end database appliances; with such an appliance, computing time is reduced to approximately 30 seconds. This version is targeted for large data sets, multiple re-analyses, or other situations in which speed and scale are crucial.

Diagnostics and presentation of results

Automated variable selection algorithms can seem to be something of a 'black box'; to counter this, we have created an extensive Web site (http://www.hdpharmacoepi.org) for making hd-PS's activities transparent. If the user requests, the algorithm can automatically upload and archive aggregated diagnostic information and anonymous summary data that is similar to the typical cohort description published in research articles. Once it is uploaded, a public link can be generated for study investigators and external reviewers to interactively browse the variables selected, review Z-bias screening reports, check other diagnostics, and compare the variables selected across studies with similar exposures and outcomes. Note that though the site can display extensive information about each analysis, no individual patient data is ever in any way transmitted, visible, or inferable.

CONCLUSION

Active safety monitoring systems will require certain decisions to be made by investigators; at the same time, meaningful scalability will require as many study elements as possible to be automated. We provide reasons and some evidence that covariate creation and selection can be accomplished effectively with an automated process such as hd-PS. We recognize the trade-offs—including variables that may be chosen because of observed associations rather than from subject matter expertise, or variables that may be included unnecessarily or even incorrectly—but in studies to date, hd-PS has made choices that provide equal or better confounding adjustment as compared to investigator-driven covariate selection, and we have no evidence that 'over-adjustments' resulting in M-bias and Z-bias are threats to validity in realistic safety surveillance settings. In general, settings of very few exposed combined with rare outcomes will remain challenging, as will settings with very strong unmeasured confounding.

DISCLOSURES

JR and SS are investigators of the Mini Sentinel project funded by the Food and Drug Administration. Mini-Sentinel is funded by the Food and Drug Administration through the Department of Health and Human Services contract number HHSF223200910006I.

In addition, this work was partially funded by research grants from the National Center for Research Resources (grant no. RC1-RR028231), the National Library of Medicine (grant no. R01-LM010213), and the National Heart Lung and Blood Institute (grant no. RC4-HL102023). Netezza, an IBM company, has donated a TwinFin 12 database appliance to the Division of Pharmacoepidemiology and Pharmacoeconomics. JR is a recipient of a career development award from Agency for Healthcare Research and Quality (K01 HS018088). He is a consultant to WHISCON. SS is Principal Investigator of the Brigham and Women's Hospital DEcIDE Center on Comparative Effectiveness Research and the DEcIDE Methods Center both funded by AHRQ. SS is paid member of the Scientific Advisory Board of HealthCore and consultant to WHISCON and Booz & Co, and he is recipient of investigator initiated grants from Pfizer and Novartis.

KEY POINTS

- To reach full scale, the Sentinel System and other drug safety surveillance systems will require automation of large parts of the evaluation process.
- We present the high-dimensional propensity score as one option for automating confounding control in safety surveillance evaluations.
- We discuss covariate selection, potential for collider bias, practical considerations, and other challenges and solutions.
- We conclude that despite certain limitations, hd-PS offers substantial advantages over non-automated alternatives in active product safety monitoring systems.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

The authors acknowledge the members of the Mini-Sentinel Methods Development Signal Evaluation Workgroup for their insights on collier bias and other variable selection challenges.

REFERENCES

1. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol* 1991; **133**(2): 144–153.
2. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol* 2003; **158**(9): 915–920.
3. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. *Pharmacoepidemiol Drug Saf* 2010; **19**(8): 858–868.
4. When Should Multi-Site Electronic Healthcare Database Surveillance Systems Use Case-Based Designs For Medical Product Safety Monitoring?: Working Group on Case-Based Approaches for the Methods Core of the Mini-Sentinel Initiative; 2011.
5. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999; **10**(1): 37–48.
6. Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. *Am J Epidemiol* 2011; **173**(12): 1404–1413.
7. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 2009; **20**(4): 512–522.
8. Rassen JA, Choudhry NK, Avorn J, Schneeweiss S. Cardiovascular outcomes and mortality in patients using clopidogrel with proton pump inhibitors after percutaneous coronary intervention or acute coronary syndrome. *Circulation* 2009; **120**(23): 2322–2329.
9. Schneeweiss S, Patrick AR, Solomon D, *et al*. The comparative safety of antidepressant agents in children regarding suicide attempts and suicides. *Pediatrics* 2010; **125**(5): 876–888.
10. Schneeweiss S, Patrick AR, Solomon D, *et al*. Variation in the risk of suicide attempts and completed suicides by antidepressant agent in adults: A propensity score-adjusted analysis of 9 years of data. *Arch Gen Psych* 2010; **67**(5): 497–506.

DOI: 10.1002/pds

11. Patorno E, Bohn RL, Wahl PM, *et al*. Anticonvulsant medications and the risk of suicide, attempted suicide, or violent death. *JAMA* 2010; **303**(14): 1401–1409.
12. Huybrechts KF, Rothman KJ, Silliman RA, Brookhart MA, Schneeweiss S. Risk of Death and Hospitalization for Major Medical Events after Initiation of Psychotropic Medications in Older Adults Admitted to Nursing Homes. *Can Med Assoc J*. 2011; **183**: E411–E419.
13. Toh S, García Rodríguez LA, Hernan MA. Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records. *Pharamcoepidemiol Drug Saf* 2011; **20**: 849–857.
14. Bombardier C, Laine L, Reicin A, *et al*. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. *N Engl J Med* 2000; **343**(21): 1520–1528, 1522 p following 1528.
15. Silverstein FE, Faich G, Goldstein JL, *et al*. Gastrointestinal toxicity with celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis: the CLASS study: A randomized controlled trial. Celecoxib Long-term Arthritis Safety Study. *JAMA* 2000; **284**(10): 1247–1255.
16. Rassen JA, Doherty M, Huang W, Schneeweiss S. HD Pharamcoepi Web Site. 2011; http://www.hdpharmacoepi.org
17. Bross IDJ. Spurious effects from an extraneous variable. *J Chronic Dis* 1966; **19**(637–47).
18. 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. *Am J Epidemiol* 2010; **172**(7): 843–854.
19. Rassen JA, Avorn J, Schneeweiss S. Multivariate-adjusted pharmacoepidemiologic analyses of confidential information pooled from multiple health care utilization databases. *Pharmacoepidemiol Drug Saf* 2010; **19**: 848–857.
20. Mahalanobis PC. On the generalized distance in statistics. *Proc Natl Inst Sci India* 1936; **12**: 49–55.
21. 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. *J Clin Epidemiol* 2009; **62**(12): 1233–1241.
22. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med* 2007; **26**(1): 20–36.
23. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997; **127**(8 Pt 2): 757–763.
24. Judkins DR, Morganstein D, Zador P, Piesse A, Barrett B, Mukhopadhyay P. Variable selection and raking in propensity scoring. *Stat Med* 2007; **26**(5): 1022–1033.
25. Weinberg CR. Toward a clearer definition of confounding. *Am J Epidemiol* 1993; **137**(1): 1–8.
26. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology* 2003; **14**(3): 300–306.
27. Pearl J. On a class of bias-amplifying variables that endanger effect estimates. Paper presented at: Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence; AUAI, Corvallis, OR, 2010.
28. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. *Med Care* 2010; **48**(6 Suppl): S114–120.
29. Liu W, Brookhart MA, Setoguchi S. Impact of collider-stratification bias (M-bias) in pharmacoepidemiologic studies: a simulaiton study. *Pharmacoepidemiol Drug Saf* 2010; **19**(S1): S212.
30. Myers J. Effects of adjusting for instrumental variables on bias and variance of effect estimates. *Am J Epidemiol* 2011; **174**: 1213–1222.
31. Pearl J. Invited Commentary: Understanding Bias Amplification. 2011; http://bayes.cs.ucla.edu/csl_papers.html. [30 June 2011].
32. Angrist JD, Imbens G, Rubin DB. Identifcation of causal effects using instrumental variables. *JASA* 1996; **94**(434): 444–455.
33. Rassen JA, Brookhart MA, Mittleman MA, Glynn RJ, Schneeweiss S. Instrumental variables I: exploiting quasi-random treatment choice to construe causal relationships. *J Clin Epidemiol* 2009; **62**(12): 1226–1232.
34. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf* 2010; **19**: 537–554.
35. Schneeweiss S, Gagne JJ, Glynn RJ, Ruhl M, Rassen JA. Assessing the comparative effectiveness of newly marketed medications: Methodological challenges and implications for drug development. *Clin Pharmacol Ther* 2011; **90**: 777–790.
36. *Pharamcoepidemiology Toolbox version 2 [computer program]*. Boston, MA, 2011.
37. Schneeweiss S, Seeger JD, Maclure M, Wang PS, Avorn J, Glynn RJ. Performance of comorbidity scores to control for confounding in epidemiologic studies using claims data. *Am J Epidemiol* 2001; **154**(9): 854–864.
38. Rassen JA, Avorn J, Schneeweiss S. Multivariate-adjusted pharmacoepidemiologic analyses of confidential information pooled from multiple health care utilization databases. *Pharmacoepidemiol Drug Saf* 2010; **19**(8): 848–857.
39. Rassen JA, Solomon DH, Curtis LH, Herrington L, Schneeweiss S. Privacy-maintaining propensity score-based pooling of multiple databases applied to a study of biologics. *Med Care* 2010; **48**(6 Suppl): S83–89.
40. Rassen JA, Glynn RJ, Rothman KJ, Setoguchi S, Schneeweiss S. Applying propensity scores estimated in a full cohort to adjust for confounding in subgroup analyses. *Pharmacoepidemiol Drug Saf* 2011. DOI: 10.1002/pds.2256. (in press)
41. Kloss S. *Propensity Score & High-Dimensional Propensity Score Methods in Observational Studies based on Administrative Data of Statutory Health Insurances*. Bremen, Universitat Bremen, 2010.
42. Rassen JA, Choudhry NK, Avorn J, Schneeweiss S. Cardiovascular outcomes and mortality in patients using clopidogrel with proton pump inhibitors after percutaneous coronary intervention or acute coronary syndrome. *Circulation* 2009; **120**(23): 2322–2329.