MINI-SENTINEL METHODS

DEVELOPMENTS, APPLICATIONS, AND METHODOLOGICAL CHALLENGES TO THE USE OF PROPENSITY SCORE MATCHING APPROACHES IN FDA'S SENTINEL PROGRAM

Prepared by: John G. Connolly, SM,[1] Judith C. Maro, PhD,[2] Shirley V. Wang, PhD,[1] Candace C. Fuller, PhD,[2] Sengwee Toh, ScD,[2] Catherine A. Panozzo, PhD,[2] Hannah Katcoff, MPH,[2] Noelle Cocoros, DSc,[2] Xu Han, PhD,[3] Meijia Zhou, MHS, [3-4] Joshua J. Gagne, PharmD, ScD[1]

Author Affiliations: 1. Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; 2. Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA; 3. Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA; 4. Center for Pharmacoepidemiology Research and Training, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

October 5, 2016

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

Sentinel

Mini-Sentinel Methods

Developments, Applications, And Methodological Challenges to The Use Of Propensity Score Matching Approaches In FDA's Sentinel Program

Table of Contents

I. ABSTRACT/EXECUTIVE SUMMARY 1

II. INTRODUCTION 1

III. NEW PROPENSITY SCORE DEVELOPMENTS AND APPLICATIONS RELEVANT TO SENTINEL 2

A. High-dimensional Propensity Scores 2
1. Development and Empirical Evaluations 2
2. Simulated Evaluations 2
B. Propensity scores Small Samples 3
C. Propensity Scores in Distributed Data Settings 4
D. Propensity Scores in Prospective Analyses 4
E. Propensity Scores in Both Distributed Settings and Prospective Analyses 4

IV. PROPENSITY SCORE APPLICATIONS IN SENTINEL 5

A. One-Time Drug Safety Assessments 5
1. Dabigatran and Intracranial or Gastrointestinal Hemorrhage, Ischemic Stroke, and Acute Myocardial Infarction 5
2. Apixaban and Intracranial or Gastrointestinal Hemorrhage and Stroke 6
3. Niacin on Intracranial or Gastrointestinal Hemorrhage and Ischemic Stroke 6
4. Levetiracetam and Agranulocytosis 6
B. Prospective Drug Safety Assessments 7
1. Rivaroxaban and Intracranial or Gastrointestinal Hemorrhage and Ischemic Stroke 7
2. Mirabegron and Stroke and Acute Myocardial Infarction 8

V. DISCUSSION 8

VI. CONCLUSION 9

VII. REFERENCES 10

VIII. TABLES AND FIGURES 11

I. ABSTRACT/EXECUTIVE SUMMARY

In the Food and Drug Administration's Sentinel program, propensity scores can be used to minimize the need for data sharing while also addressing potential confounding in medical product safety surveillance activities. Recent developments in propensity score methodology led to the creation of a propensity score matching tool that was first added to the Sentinel program in 2013. To date, the propensity score matching methods utilized by the tool have been validated on several known positive and negative drug-outcome associations. Within Sentinel, the propensity score matching tool has also been used to successfully conduct post-approval safety surveillance of newly approved or older medications in both static and dynamic data sources. These experiences have highlighted areas of improvement for future versions of the tool, as well as unresolved questions regarding the use of propensity score methods in distributed and prospective data environments which include: 1) the optimal approach for estimating propensity score models with many covariates within small Data Partners where new users are scarce, 2) the optimal group in which to estimate the propensity score in sequential data, and 3) whether it is better to lock data on matched sets and outcomes from previous monitoring periods as new data accumulates, or instead allow for changes to past data with each data update.

II. INTRODUCTION

Propensity scores (PSs) are an important confounding adjustment technique for active post-approval medical product safety surveillance systems such as the US Food and Drug Administration's (FDA's) Sentinel program. The PS is the probability of treatment assignment conditional on observed baseline characteristics, and can balance measured patient characteristics in much the same way that randomization balances both observed and unobserved potential confounders.[1] PSs have typically been used in retrospective analyses of single databases. In contrast, Sentinel operates a distributed database in which identical analyses are performed separately within each participating Data Partner (DP) and then aggregated into a final effect measure. Some Sentinel analyses are also sequential, and the Sentinel initiative utilizes scalable and semi-automated tools so that analyses can be performed quickly to provide timely safety to FDA.

By reducing a vector of covariates into a single number, PSs facilitate confounding adjustment for a potentially large number of confounding variables without being limited by the number of outcome events. Recent methodological advances have made the PS a particularly useful method in the distributed data setting when the number of outcome events is often very small within each individual DP, where analyses are performed. Furthermore, PSs enable multivariable adjusted analyses with minimal data sharing in the distributed data setting, where it is important to limit sharing of patient level data due to privacy concerns.

For these reasons, PSs have been incorporated into the routine querying framework for the FDA's Active Risk Identification and Analysis (ARIA) system. A component of Sentinel, ARIA relies on existing, customizable analytic tools and quality checked administrative and electronic health record data in the Sentinel Distributed Database to perform post-approval safety surveillance of FDA-regulated medical products. In particular, the Propensity Score Matching (PSM) tool combines a new user, active comparator cohort design with PS matching, which has many strengths, including that it ensures accurate temporality of exposure and outcome, prevents common biases in observational analyses, such as baseline selection bias and immortal person-time bias, and reduces confounding. The PSM tool is

compatible with the Sentinel Common Data Model (CDM), a standardized structure. Participating DPs regularly convert their data to the Sentinel CDM in order to make them available for assessment. The PSM tool was designed to perform both retrospective and prospective analyses.

Since 2013, the Sentinel program has been evaluating and using the PSM tool to assess the safety of various drugs. As part of the development and enhancement process, the tool has also been used to replicate known associations in the Sentinel Distributed Database. The purpose of this report is to describe the development, application, and performance of PS-based approaches used in Sentinel, with a particular focus on the challenges and successes of these approaches in the distributed analysis setting. Some evaluations were performed sequentially and those are reviewed for their sequential elements. First, we review methodological advances from the literature relevant to the use of PSs for post-approval surveillance such as the high dimensional propensity score (hd-PS) and plasmode simulation. Next, we summarize applications of the PSM tool within the Sentinel program including test cases where the drug-outcome association was known as well as one-time and sequential safety analyses.

III. NEW PROPENSITY SCORE DEVELOPMENTS AND APPLICATIONS RELEVANT TO SENTINEL

A. HIGH-DIMENSIONAL PROPENSITY SCORES

1. Development and Empirical Evaluations

Although the PSM tool streamlines the design and analysis of safety assessments, it does not automate decision-making about various important clinical, epidemiologic, and regulatory inputs, such as the medical product exposure of interest, the health outcome of interest, and the most appropriate comparator product. Investigators must also specify the potential confounders to be included in an analysis in order to generate valid safety evidence. The hd-PS algorithm was proposed as a possible solution for automating the selection of potential confounders empirically.[2] The principle behind hd-PS is that measured variables (e.g., oxygen tank usage) can serve as proxies for unmeasured confounders (e.g., frailty) and certain variables in the data may behave as confounders even though the investigator did not think to measure and adjust for them. By including these variables in a PS, one can partly adjust for variables that are not directly measured and can realize greater confounding control than by relying solely on investigator-specified variables. In order to identify covariates for inclusion, hd-PS automatically identifies and ranks variables based on their potential to cause confounding, as measured by a variable's respective associations with exposure and/or outcome. Empirical assessments suggest that including only variables identified by the hd-PS algorithm seems to perform as well as including a combination of investigator-specified and hdPS-identified variables in terms of confounding adjustment.[2] The hd-PS algorithm is available as an option in the PSM tool.

2. Simulated Evaluations

Simulation studies have shown that automated variable selection, as implemented in hd-PS, will rarely cause problems with respect to bias in a measure of association.[3,4] hd-PS has also been more fully evaluated using a plasmode simulation framework, in which simulated outcomes are injected into real electronic healthcare data to provide a more realistic model of actual data than fully synthetic simulation approaches. This simulation platform allows for previously impossible evaluations of hd-PS in simulated settings given the complexity of the correlation structure of pre-treatment variables.[5] The hd-

PS algorithm was evaluated in the context of an example comparing new users of high versus low-intensity statins for the prevention of cardiovascular events.[5] PSs were estimated using investigator-specified covariates only as well as two sets of hd-PS-identified variables, one where covariates were ranked by the strength of their association with exposure (i.e., exposure-based), and one where they were ranked by their associations with both exposure and outcome (i.e., bias-based). Effect estimates after adjusting for PS deciles showed that hd-PS removed more bias than investigator-specified variable selection in the presence of a non-null treatment effect. Another important finding was that the bias reduction from an exposure-based hd-PS was equivalent to the bias-based hd-PS in both null and non-null scenarios. This finding is particularly reassuring when using hd-PS in the early post-approval period where there are likely to be few outcomes.

B. PROPENSITY SCORES SMALL SAMPLES

In the early post-approval period, it is in the best interest of patients and regulators to detect previously unknown safety issues as soon as possible. However, in the early post-approval period there will typically be few outcomes and few users of the new medication relative to the comparator. The performance of hd-PS in settings where there are few exposed patients or outcome events was evaluated by re-sampling from four new user cohorts identified in administrative claims data to create increasingly smaller cohorts.[6] The drug-outcome pairs were COX-2 inhibitors and gastrointestinal complications, statins and all-cause mortality, statins and acute myocardial infarction/non-cancer mortality, and selective serotonin reuptake inhibitors (SSRIs) and suicidal acts. In these small cohorts, the authors examined the performance of hd-PS and sought to define the optimal number of empirically identified covariates to include. When there were greater than 50 exposed patients with an event, the bias-based hd-PS algorithm performed well in terms of confounding adjustment. When there were fewer than 50 exposed events, ranking covariates for inclusion based only on their association with exposure improved the algorithm's performance. When there were between 25-49 exposed events, adding a zero-cell correction improved confounding adjustment, but using the exposure-only ranking method provided more reliable results across all examples. The PSM tool enables investigators to specify which hd-PS algorithm to use (e.g., bias-based or exposure-based) and whether to use a zero-cell correction.

The aforementioned issues with small samples may also arise when performing subgroup analyses; for example, when investigating whether a drug is more likely to cause an adverse event in older versus younger individuals. A later study investigated whether a PS estimated in the full cohort remained valid for subgroup analyses, or whether a subgroup-specific PS would perform better.[7] In order for the full cohort PS to remain valid, theory states that the subgroup variable must be included in the full cohort PS and that the original population and subgroup need to be large enough for large sample theory to hold. Using two empirical new user cohorts (COX-2 inhibitors and gastrointestinal toxicity, antipsychotics and mortality) and one simulation study, the authors investigated PS performance after 1:1 matching in situations with infrequent outcomes, small subgroups, strong confounding, and covariate interactions. They found the full cohort PS provided adequate confounding adjustment within subgroups except when subgroups contained fewer than 500 patients or there were few exposed patients with events. In general, misspecification of the PS model was more important to PS performance than whether the PS was estimated in the full cohort or a subgroup. Based on these findings, the PSM tool enables subgroup analyses on the basis of any pre-specified variable by using the PS values from the full cohort to re-match individuals within each level of the subgroup variable.

C. PROPENSITY SCORES IN DISTRIBUTED DATA SETTINGS

Compared with analyses in single databases, performing analyses in multiple databases improves statistical power, which can enable more rapid identification of potential drug safety issues. A challenge of the distributed data setting is maintaining confounding control while simultaneously minimizing or eliminating sharing of patient-level data. Two recent studies compared various methods for achieving this goal, which included comparing full data sharing, cell-aggregated sharing, distributed regression analysis, meta-analysis, and PS-based methods.[8,9] Both evaluations concluded that PS methods provided the best flexibility, confidentiality, and analytic integrity. Using such methods, individual DPs estimate a PS within their own data and share only the PS, along with a randomly generated patient identifier and relevant subgroup indicators, with a central analysis hub like the Sentinel Operations Center (SOC). The feasibility of this PS-sharing approach was demonstrated in a cohort of new users of clopidogrel and PPIs versus new users of clopidogrel alone monitored for risk of myocardial infarction. Data were shared among four DPs without compromising patient privacy by using PSs.[8]

D. PROPENSITY SCORES IN PROSPECTIVE ANALYSES

Recent studies have investigated the use of PSs for conducting prospective, sequential safety monitoring in both single databases and distributed data settings. One such study developed an approach to prospective, targeted safety monitoring and evaluated it using three drug-outcome pairs with known safety issues: paroxitene and severe upper gastrointestinal bleed, lisinopril and angioedema, and ciprofloxacin and Achilles tendon rupture.[10] The authors divided a single dataset into sequential monitoring periods to mimic the prospective accumulation of data, and 1:1 matched new users of the study drugs within each monitoring period on a PS estimated only among those new users. Two alerting algorithms were compared: a maximized sequential probability ratio test (maxSPRT[11]) and a method which generated an alert if the effect estimate exceeded a predefined threshold for more than 3 consecutive monitoring periods. There were no corrections made for multiple testing in the second algorithm. Alerts were generated by at least one of the alerting algorithms in each example over the follow up period, demonstrating the feasibility of such a PS-matched prospective monitoring system to quickly identify adverse events. Another option for estimating PSs in the prospective setting is to include all patients in each new PS model each time the data are updated prospectively. This is the approach used in the PSM tool. Using this approach, patients from prior monitoring periods can be re-matched with each new PS or they can remain in their original matched pairs or sets. It is important to note that while these assessments emulated prospective analyses, the dataset was static and therefore avoided issues related to changing data over time that can occur when using dynamic datasets in the Sentinel Distributed Database.

E. PROPENSITY SCORES IN BOTH DISTRIBUTED SETTINGS AND PROSPECTIVE ANALYSES

One of the concerns regarding automated safety surveillance systems in the past, such as the Centers for Disease Control and Prevention's (CDC's) Vaccine Safety Datalink (VSD), is that there will be too many false positives or false negatives.[12] To address these concerns, a semi-automated safety monitoring approach was utilized to assess the safety of three medications across three claims databases.[13] Data from each database were divided into three-month intervals to mimic prospective accumulation of data, and new users of the three study drugs were 1:1 PS matched to new users of comparator products within each interval. The three drug-outcome relationships were rosuvastatin and rhabdomyolysis, rosuvastatin and diabetes mellitus, and telithromycin and hepatotoxicity. As expected based on prior research, none of these three examples generated a safety alert, demonstrating the

system's robustness against false positives. By sharing only PSs with a central data analysis center, confounding was mitigated without compromising patient privacy.

A subsequent paper described in more detail the prototype for the PSM tool and the steps required to conduct prospective semi-automated surveillance using PS matching in distributed databases.[14] The PSM tool has several functions and includes each of the features noted above. When tested on retrospective data for nine drug-outcome pairs, which included positive and negative controls, the tool generated results consistent in direction and magnitude with expectations for each. Alerts were generated for all positive controls when power was adequate, which included the three known positive associations described above. These findings further demonstrate the ability of the system to avoid generating false positives and negatives.

The PSM prototype was also applied to a prospective safety monitoring scenario in which data accumulated in real time to evaluate the risk of hemorrhagic and ischemic events in patients taking prasugrel versus clopidogrel during the first two years of prasugrel's market availability in the US.[15] Using the prototype, patients were matched on hd-PS and followed for outcomes using dynamic data that were updated on a bi-monthly basis. In general, results were similar to those from randomized trials. While the first monitoring period had to be lengthened in order to accumulate enough users to validly estimate the hd-PS, the tool performed well and demonstrated the feasibility of this type of analysis in a prospective, distributed database setting like Sentinel.

IV. PROPENSITY SCORE APPLICATIONS IN SENTINEL

Below we summarize the applications of the PSM tool in the Sentinel environment, challenges encountered, and lessons learned from each assessment. These are grouped based on whether the analyses were single one-time assessments or prospective sequential assessments. **Table 1** includes a tabular summary of each application.

A. ONE-TIME DRUG SAFETY ASSESSMENTS

1. Dabigatran and Intracranial or Gastrointestinal Hemorrhage, Ischemic Stroke, and Acute Myocardial Infarction

The analysis compared new users of dabigatran to new users of warfarin among patients who had a diagnosis of atrial fibrillation and lacked a diagnosis of valve disease or replacement, deep vein thrombosis or pulmonary embolism, dialysis, or joint replacement in the 183-day lookback period. As a sensitivity analysis, the lookback period was extended to 365 days for each of the four outcomes. The four outcomes of interest were intracranial hemorrhage, gastrointestinal hemorrhage, ischemic stroke, and acute myocardial infarction that occurred as the inpatient primary diagnosis. Intracranial hemorrhage could also occur as a secondary inpatient diagnosis. Patients were 1:1 matched on an investigator-specified PS within calipers of 0.025 and 0.050, and 95% of dabigatran users were matched using both calipers. The query period spanned from November 1st, 2010 to December 31, 2013.

Separate runs were executed among those aged 21 years or older and those aged 65 years or older. This was done because of a specific interest in examining results separately for older males and females. The PSM tool does not currently allow for subgroup analyses within subgroups defined simultaneously by two or more variables. However, restricting the analysis on one of the variables (e.g., those aged 65 years or older) and conducting a subgroup analysis by the other variable (e.g., sex) allows this to be done. At some smaller DPs, the PS models converged in the 21+ analyses, but failed to converge in

analyses restricted to these 65+ grouping due to small numbers of patients for some outcomes. Conducting separate runs within males and females and then performing subgroup analyses within those groups may have avoided the non-convergence issues. Nevertheless, future enhancements should modify the PSM tool to enable more streamlined analyses within subgroups defined by more than one variable.

2. Apixaban and Intracranial or Gastrointestinal Hemorrhage and Stroke

This assessment compared new users of apixaban to new users of warfarin among patients with a diagnosis of atrial fibrillation and no evidence of renal transplant or dialysis, valve disease or replacement, or knee or hip replacement. The outcomes of interest were intracranial hemorrhage, gastrointestinal hemorrhage, and stroke occurring as the inpatient primary diagnosis. New users of apixaban were 1:1 matched on an investigator-specified PS to new users of warfarin within calipers of 0.025 and 0.05, with over 99% of apixaban initiators matched using both calipers The query was sent to the four largest DPs, and the query period spanned from February 1, 2013 through May 31, 2015.

Analyses were conducted both with and without stratifying the outcome model on the matched pair. Conditioning on the analysis on the matched pair (i.e., using a stratified outcome model) is not necessary for controlling for baseline confounding after fixed-ratio PS matching; however, stratified analyses were performed in order to satisfy the conditions of the maxSPRT sequential reporting algorithm which required equal follow-up time between exposure groups. Stratified analyses achieve this by censoring the entire matched set upon loss to follow-up of any member of that set. Although confidence intervals were wide because few outcomes were included in this assessment, it was evident that stratifying on the matched pair resulted in fewer informative events in the analysis and even wider confidence intervals as compared to not stratifying the model. It was decided that, in general, conditional analyses when fixed ratio matching are not preferable in the Sentinel context because of the lower statistical efficiency. Unlike fixed ratio matching, when variable ratio matching a stratified or conditional Cox model is necessary. However, this approach may lead to some events being non-informative within sets. Alternative approaches for analyzing variable ratio matched data are stratifying on ratio size or weighting by the size of the matched set.

3. Niacin on Intracranial or Gastrointestinal Hemorrhage and Ischemic Stroke

This analysis compared new users of niacin or niacin+statin combination drugs to new users of fenofibrate. The outcomes of interest were major gastrointestinal bleeding, intracranial hemorrhage, and ischemic stroke occurring as a primary inpatient diagnosis. The query period was from January 1, 2007 through May 31, 2013 and the request was sent to and returned by the four largest DPs. In each analysis, patients were 1:1 matched on an investigator-specified propensity score using a caliper of 0.05 and over 90% of niacin initiators were matched. No challenges with the use of tool were identified in this example, in part because this query was sent to the largest Data Partners who provided ample new users to reliably estimate the PS models, but did not have such extremely large cohorts that DP specific memory issues occurred.

4. Levetiracetam and Agranulocytosis

The risk of agranulocytosis events was compared among new users of oral levetiracetam who did not have a diagnosis of malignant cancer or other certain chronic conditions using two comparator groups: new users of lamotrigine and new users of topiramate. Sensitivity analyses were conducted for both versions where agranulocytosis need not occur as the inpatient primary diagnosis, but could instead occur as a non-secondary inpatient diagnosis. The query was distributed to 17 DPs on September 26,

2014, and the query period for the request was January 1, 2000 to October 31, 2013. Results were returned from 10 DPs where the program ran without error and the PS models converged. Patients were 1:1 matched on an investigator-specified PS using calipers of 0.01 and 0.05. There were 12 user-defined covariates.

Of the seven sites where the program failed to run, four had errors in their logs, the PS model failed to converge at two sites, and convergence errors occurred at one site. The standard examination of DP-specific PS histograms performed in all examples showed that, prior to matching, the exposure and both comparator groups were substantially different from one another at some sites. Across DPs returning results, approximately 82% of levetiracetam users were matched to lamotrigine users, and 81% of levetiracetam users were matched to topiramate users. After matching, there were residual imbalances in the age of the cohorts. SOC epidemiologists expressed concerns about potential channeling bias – e.g., females at high risk for agranulocytosis should not have been given lamotrigine in the first place, and so lamotrigine was preferentially prescribed to low-risk users.

B. PROSPECTIVE DRUG SAFETY ASSESSMENTS

1. Rivaroxaban and Intracranial or Gastrointestinal Hemorrhage and Ischemic Stroke

This ongoing assessment compares new users of rivaroxaban to new users of warfarin among patients with a diagnosis of atrial fibrillation and no evidence of renal transplant or dialysis, valve disease or replacement, or knee or hip replacement in a 183-day lookback period. Individuals are followed over their initial treatment episode for the outcomes of ischemic stroke, intracranial hemorrhage and gastrointestinal bleeding. Patients are variable-ratio matched on both an investigator-specified and hd-PS using an upper limit of 10. The query period began on November 1, 2011 and includes the four largest DPs. Sequential testing is performed using the Wald statistic with a two-sided test.

Following the first look, the Workgroup identified that the tool has limited ability to diagnose differential loss-to-follow-up with available outputs. Future enhancements to the PSM tool will involve the development of additional diagnostic outputs, including Kaplan-Meier survival curves and plots depicting distributions of follow-up time for each exposure group.

The Workgroup also encountered problems due to the dynamic nature of the prospectively accumulating data. Because Sentinel uses recently collected data, the data are liable to change over time as delayed claims enter the databases or individuals are removed from databases for administrative reasons. Patient identifiers can also change over time, which would preclude following patients from one data update to the next. The original PSM prototype included options for addressing each of these issues. As a first option, the tool enables the user to preserve matched pairs or sets from prior monitoring periods as new data accrue over time. This ensures that, even in the face of changes to the underlying data, cohorts identified and included in the assessment from prior monitoring periods do not change. As a second option, the tool enables re-matching of patients from each monitoring period, which allows investigators to include patients across multiple monitoring periods even if their identifiers change. Because an optimal nearest-neighbor matching process is used, this second option ensures that the PS estimation and matching process produces the same matched pairs or sets provided that previously analyzed data do not change across looks. However, any changes to data that were included in the analysis at a prior look, including the addition or subtraction of new users in a monitoring period, can induce changes to the PS estimation model for that period. Slight changes to the PS can shift the optimal matches for individuals. The rivaroxaban assessment initially used the second approach but, in light of findings that most matches changed due to the changing data, the first approach is being used in

future analyses. The optimal solution for simultaneously addressing both data limitations is not known and is the subject of ongoing work.

2. Mirabegron and Stroke and Acute Myocardial Infarction

This analysis compared new users of mirabegron versus new users of oxybutinin. The outcomes of interest were stroke and acute myocardial infarction. The primary version of the analysis required patients to be naïve to prescription overactive bladder treatment prior to initiation of mirabegron or oxybutinin, while a secondary version required prior use of another overactive bladder medication. The queries were executed at four DPs, and the query period was from November 1, 2012 through December 31, 2013. After the first look, it was determined that uptake of mirabegron was too slow to allow for formal sequential analyses. No additional challenges were identified in the context of this assessment.

V. DISCUSSION

To date, the methods utilized by the PSM tool have performed well both inside and outside of the Sentinel program. The successful performance of the tool in such settings allowed investigators to confidently apply the PSM tool to true safety analyses despite the known limitations of PS matching.

The applications described above have identified certain challenges when using the PSM tool. In some Sentinel applications with distributed data there were issues with PS model convergence at smaller DPs, especially when using hd-PS, because there weren't enough new users to estimate PS models with many covariates. A challenge unique to dynamic datasets, such as in the rivaroxaban example, is the aforementioned problem with matched sets and other patient data changing between monitoring periods as new data accumulated and PSs changed over time. Another challenge to sequential analysis is slow uptake of the drug of interest, as in the mirabegron example, which can preclude sequential analyses even in large distributed data settings like Sentinel. In addition to difficulties fitting large PS models in such settings, when there are few new users it is important to monitor overlap of the PS distributions as this will determine what proportion of new users of the drug of interest can be matched.

Steps have already been taken to correct some of the issues identified above. For example, in prospective analyses, investigators now exercise the option to lock data from prior matching periods, including matched set information, which prevents changes to past data as new data accumulates.

However, other questions regarding the PSM tool remain unresolved. The best approach for estimating PSs at small DPs with too few exposed patients to estimate large PS models is still unknown. Also unresolved is the best way to estimate and match on the PS in sequential data. While the PSM tool currently re-estimates a PS in the entire cohort after each monitoring period and preserves the original matched sets, investigators have alternatively re-estimated PSs and re-matched among only new users within a given monitoring period. A related issue is whether, in prospective analyses, it is better to lock data from prior monitoring periods, including matched sets, or allow it to change with each data update. For example, due to delays in claims processing a patient who had an outcome in the first monitoring period may not be classified as such until the end of the second monitoring period. Current practice is to preserve past covariate data but allow outcome data to be changed, though it is uncertain if this is the optimal approach.

There are several desired future enhancements to the PS matching tool regarding model convergence, data characterization, and reporting. In terms of defining and creating covariates, desired enhancements include: the ability to choose unique covariate assessment windows for each covariate and to choose

whether to include the index date in the covariate assessment window separately for each covariate, the ability to choose granularity for empirically defined covariates (e.g. 3, 4, or 5 digit ICD-9 codes; generic name vs. drug class vs. other for empirically identified drug covariates), the ability to specify functional form (i.e. continuous vs. binary) of pre-defined covariates in PS estimation model, an increased flexibility in covariate definitions which allows complex code definitions (e.g. diagnosis code plus dispensing), and the ability to define subgroups to be defined by variables not included in the PS model.

Desired enhancements to the analytical capabilities of the tool include: more streamlined analyses within subgroups defined by more than one variable, the ability to choose the maximum follow-up length for AT analyses, the ability to perform PS adjustment via inverse-probability weighting or regression in addition to matching, the ability to stratify by follow-up time to explore time varying hazards when 1:1 matching, the ability to compare multiple exposure groups through n-wise comparisons, the ability to adjust for covariates in the outcome Cox models that were not adequately adjusted for by PS matching (requires patient level data), no longer forcing any covariates to be included in the PS models, changing the default variable matching ratio to 1:4 from the current 1:10, allow the matching ratio to be specified by the user through a macro, the ability to perform variable and fixed ratio PS matching with a varying cap (2-3 for fixed or 5-10 for variable) as opposed to the current 1:1 fixed or 1:10 variable ratio matching procedures, and diagnostics to investigate specific reasons for model non-convergence which can be distributed and aid in troubleshooting.

Desired enhancements to the reporting capabilities of the tool include: the inclusion of information (follow-up time with and without censoring, total days supplied, duration of initial treatment, etc.) on exposure duration in each group when performing ITT analyses, information on the number of informative events in each exposure group prior to censoring in ITT analyses, display of the number and percentage of patients dropped in the adjustment macro for each censoring reason in the attrition table, information on the strength of association between exposure and individual covariates included in the PS, Table 1s for newly identified patients at each refresh in addition to cumulative Table 1s, subgroup specific Table 1s, output of forest plots in standard reports to display site-specific estimates, histograms for the number of controls in each matched set for evaluating variable ratio matched analyses, and display of the number of matched controls as a function of the PS.

VI. CONCLUSION

Recent developments in PS methodology have led to the creation of the PSM tool used in the FDA's Sentinel program. This tool enables investigators to control for confounding in post-approval safety monitoring conducted in distributed data networks without compromising patient privacy. The PSM tool has already been successfully applied to multiple safety assessments using both static and dynamic databases. Future investigations into the use of PS matching methods in distributed and dynamic databases should seek to address the challenges identified in this manuscript in order to improve post-approval safety monitoring in the United States.

VII. REFERENCES

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41-55.
2. 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;20:512-22.
3. Myers JA, Rassen JA, Gagne JJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. American Journal of Epidemiology 2011;174:1213-22.
4. Liu W, Brookhart MA, Schneeweiss S, Mi X, Setoguchi S. Implications of M bias in epidemiologic studies: a simulation study. American Journal of Epidemiology 2012;176:938-48.
5. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. Computational Statistics & Data Analysis 2014;72:219-26.
6. Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. American Journal of Epidemiology 2011;173:1404-13.
7. 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. Pharmacoepidemiology and Drug Safety 2012;21:697-709.
8. Rassen JA, Solomon DH, Curtis JR, Herrinton L, Schneeweiss S. Privacy-maintaining propensity score-based pooling of multiple databases applied to a study of biologics. Medical Care 2010;48:S83-9.
9. Toh S, Gagne JJ, Rassen JA, Fireman BH, Kulldorff M, Brown JS. Confounding adjustment in comparative effectiveness research conducted within distributed research networks. Medical Care 2013;51:S4-10.
10. Wahl PM, Gagne JJ, Wasser TE, et al. Early steps in the development of a claims-based targeted healthcare safety monitoring system and application to three empirical examples. Drug Safety 2012;35:407-16.
11. Kulldorff M DR, Kolczak M, Lewis E, Lieu TA, Platt R. A Maximized Sequential Probability Ratio Test for Drug and Vaccine Safety. Sequential Analysis 2011;1:58-78.
12. Yih WK, Kulldorff M, Fireman BH, et al. Active surveillance for adverse events: the experience of the Vaccine Safety Datalink project. Pediatrics 2011;127 Suppl 1:S54-64.
13. Gagne JJ, Glynn RJ, Rassen JA, et al. Active safety monitoring of newly marketed medications in a distributed data network: application of a semi-automated monitoring system. Clinical Pharmacology and Therapeutics 2012;92:80-6.
14. Gagne JJ, Wang SV, Rassen JA, Schneeweiss S. A modular, prospective, semi-automated drug safety monitoring system for use in a distributed data environment. Pharmacoepidemiology and Drug Safety 2014;23:619-27.
15. Gagne JJ, Rassen JA, Choudhry NK, et al. Near-real-time monitoring of new drugs: an application comparing prasugrel versus clopidogrel. Drug Safety 2014;37:151-61.
16. Toh S, Reichman ME, Houstoun M, et al. Comparative risk for angioedema associated with the use of drugs that target the renin-angiotensin-aldosterone system. Archives of Internal Medicine 2012;172:1582-9.
17. Kuntz JL, Chrischilles EA, Pendergast JF, Herwaldt LA, Polgreen PM. Incidence of and risk factors for community-associated Clostridium difficile infection: a nested case-control study. BMC Infectious Diseases 2011;11:194.

VIII. TABLES AND FIGURES

Table 1. Tabular Summary of Queries Using the PSM Tool within Sentinel

Query	Outcomes	#DPs Returned/ Sent	Query Period[1]
One-time drug safety assessments			
Dabigatran vs. warfarin	ICH GI Bleed Isch. Stroke AMI	4/4	11/01/2010-12/31/2013
Apixaban vs. warfarin	GI Bleed ICH Stroke	4/4	02/01/2013-05/31/2015
Niacin vs. fenofibrate	GI Bleed ICH Stroke	4/4	01/01/2007-05/31/2013
Levetiracetam vs. lamotrigine/topiramate	Agranulocytosis	10/17	01/01/2000-10/31/2013
Prospective drug safety assessments			
Rivaroxaban vs. warfarin	Isch. Stroke ICH GI Bleed	4/4	11/01/2011-Present
Mirabegron vs. oxybutinin	Stroke AMI	4/4	11/01/2012-12/31/2013

AMI - Acute Myocardial Infarction; DP – Data Partner; GI - Gastrointestinal; ICH - Intracranial hemorrhage; Isch. Stroke – Ischemic Stroke

[1]Query period represents available data across all Data Partners included in that query