January 2016

A Review Of Propensity Score Use In Drug Post-Marketing Safety Surveillance

Motao Sun
Yale University, motao.sun@yale.edu

A Review of Propensity Score Use in Drug Post-Marketing Safety Surveillance

By

Motao Sun

Dr. Robert Makuch, Advisor

Dr. Zuoheng Wang, Second Reader

A thesis submitted in partial fulfillment

of the requirements for the Degree of Master of Public Health

in Chronic Disease Epidemiology & Regulatory Affairs Concentration

Yale School of Public Health

New Haven, Connecticut

Abstract

Drug post marketing safety surveillance is a very important part of pharmacovigilance, which is overseen by the Food and Drug Administration (FDA) and other regulatory agencies throughout the world. These studies are generally performed by pharmaceutical sponsors and academicians, although it is occasionally the case that a regulatory agency will either fund the research or have large databases which is analyzed by the regulators themselves. One example of agency surveillance is the FDA Adverse Event Reporting System (AERS). Generally, preapproval studies only involve several hundred to several thousand patients, so all possible side effects of a drug cannot be explored thoroughly because it is not possible to have a high probability of detecting rare, but important, adverse events. In addition, the population of patients in pre-approval studies are more restricted and better controlled than the wider population of patients who take a compound once it has been approved.. In order to monitor adverse events and serious/severe adverse reactions, FDA maintains a system of post-marketing surveillance programs to identify adverse events that did not appear during the drug approval process[1].

In this thesis, we will examine some regulatory/analytic issues that arise in the evaluation of pharmaco-epidemiologic data analysis using propensity score analysis from drug post-marketing safety surveillance. It is my desire to show the rationale and preference of using propensity score in post-marketing safety surveillance studies. We also examined several independent cases that used propensity score inappropriately, and concluded the FDA's comments on them. We propose some advices on propensity score use in the future statistical analysis for reference.

Keywords

Propensity Score; FDA; Regulatory; Post-Marketing Surveillance

Introduction and Background

Post-Marketing Drug Safety Surveillance

After a new drug gets its approval from FDA, the regulatory requirements become different and change from those applied during the approval process. Figure 1 shows some aspects of the changes that occur over the life-cycle of a drug that can exceed 15+ years. In initial human studies (Phase 1), safety information is acquired as part of the evaluation of various dosing regimens and dose-finding. This is used to determine whether the drug should be developed further, and brought into phase 2 testing where both safety and efficacy are evaluated. Further safety and efficacy assessments are brought into the evaluation of risk: benefit, and phase 3 testing proceeds if this risk: benefit profile is deemed desirable based on other drug competitors. If phase 3 testing is successful and the drug is approved, then risk: benefit has been established. Now, the drug enters the post-approval phase, and safety becomes important as a stand-alone issue. It is of interest to determine the safety in a broader population of patients than used in the pre-approval stages. Post-approval patients may be older, sicker, and healthier, have more or fewer co-morbidities, use more or fewer concomitant medications, and other factors that may put them at increased risk of safety compared to a relatively narrowly-defined population used during the approval process.

The major responsibilities facing pharmaceutical companies as well as FDA are similar in many aspects throughout the entire lifecycle of the drug. Part of FDA's mission is to "assure that patients and providers have timely and continued access to safe and effective and high quality drugs," and to "facilitate drug innovation."[2] Just as FDA must ensure that the drug is developed

and produced according to accepted standards, the company must also provide similar assurances when the product is marketed to the general public.[3]

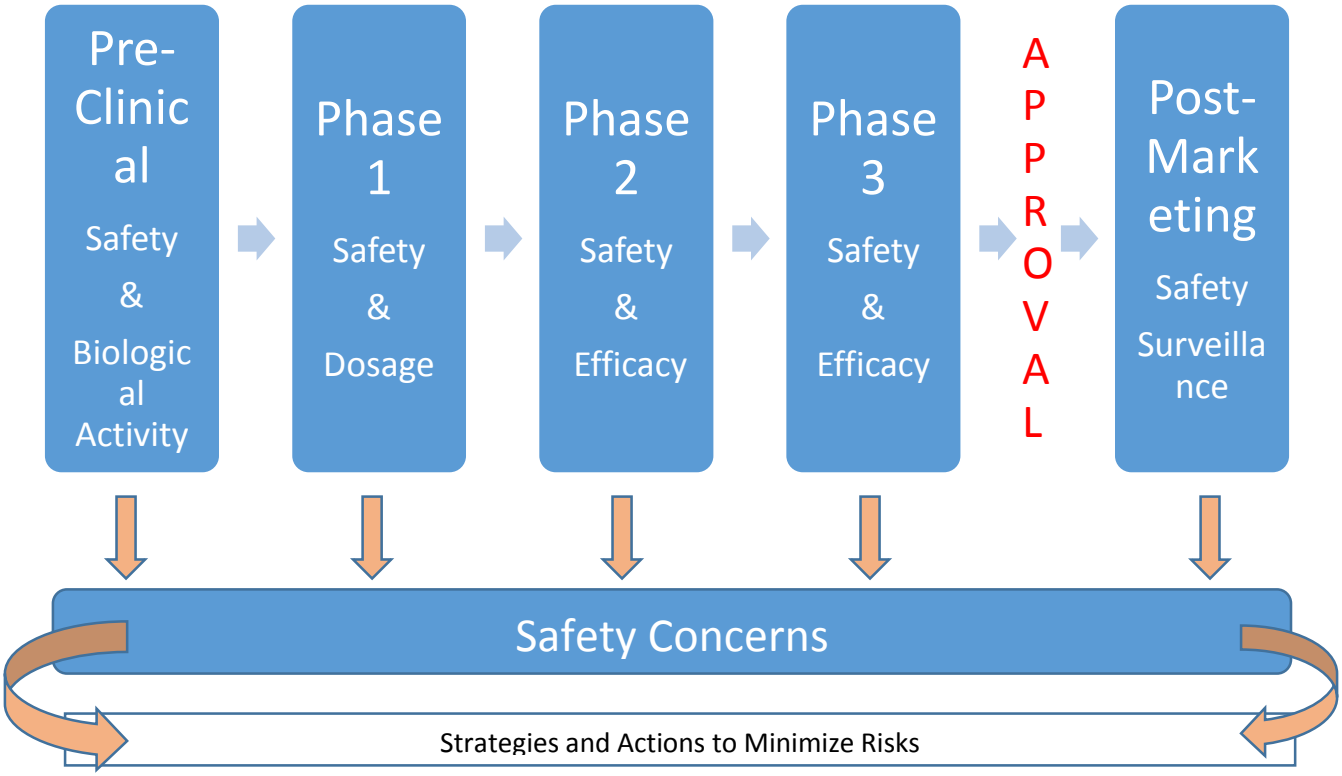


Figure 1: Safety in the Lifecycle of FDA-regulated Products

Generally, post marketing surveillance includes three types: spontaneous/voluntary reporting of cases, post-marketing studies (both voluntary and required), and active surveillance. The related pharmacy-epidemiology studies can be based on large databases collected from a variety of sources. These sources may include spontaneous clinical reporting, and assembling of data prospectively gathered from newly designed studies. Furthermore, there is also assembly of data from existing data sources such as insurance databases that are used for a variety of purposes such as billing. There are several main channels for the cases reporting: 1) FDA Med Watch system, 2) the manufacturers' required regulatory reporting, insurance databases, and 3) large governmental databases such as Medicare/Medicaid. The manufacturer/insurance/

governmental reporting accounts for roughly 95% of all reports, and Med Watch accounts for the remaining 5%.[4]

The FDA adverse event database is called the Adverse Event Reporting System (AERS), which contains information on adverse event and medication error reports submitted to FDA by the industry, physicians, lawyers, and patients. The AERS supports a major part of the FDA's post-marketing safety surveillance program for drug and therapeutic biologic products. The AERS database is a useful tool for FDA to look for new safety concerns, evaluation of manufacturer's compliance to reporting regulations, and responding to outside requests for information.[5] The database contains over 7 million reports since 1969, and nearly 1 million new reports each year recently. [6]

Since some of the adverse events classifications in the post-market phase are relative rare, observational and other types of epidemiologic/clinical studies are commonly used in this phase compared with use of prospectively randomized, controlled clinical trials in the pre-approval phase. Observational/epidemiologic studies have the advantage of: 1) being relatively inexpensive compared to randomized trials, and 2) having access to people in a broader environment of 'real life' situations.. What is more, most of the post-market safety surveillance will collect a large amount of cases, which are generally far more than what one would obtain in prospectively randomized clinical trials, so the power of the observational studies are greater and usually acceptable for making an informed scientifically valid decision. However, observational studies may be subject to several biases meanwhile, especially selection bias.[7] Channeling bias is another problem, since physicians may 'channel' or direct patients to a certain therapy more often based on their patients' personal characteristics. Furthermore, such studies may have limitation in determining causation. Nevertheless, they contribute to the totality of evidence and

thereby provide important information that can be used to evaluate the possibility of safety concerns.

Selection bias is an important concern in observational studies, especially in post-marketing observational studies where they may form the basis for possibly removing a drug from the marketplace. Often, only one study is done and so the usual requirement of scientific replication is ignored. This is due to a number of reasons, including that one study is very large and relatively lengthy. In these studies, most data are collected from clinical cases from hospitals, clinics and healthcare professionals, in which decisions of using any specific drugs are made by local physicians. The selection of any particular drug is complex, and based on their perceptions of the risks and benefits for particular patients.[8] As a result, substantial selection and channeling bias would be introduced since the allocation of treatments will be imbalanced in risk factors as well as demographics between patients given the drug of interest and those given an alternative or no treatment. That is to say, based on physicians' previous clinic experiences, a certain group of patients who share certain kinds of characteristics are more likely to receive a specific treatment because they think it will have better risk: benefit on these patients. These imbalances in patient characteristics will directly affect safety outcomes when higher risk patients receive the drug.[9] This will bias observational studies to the different directions.

The question for the analyst of safety data becomes, how these factors can be adjusted in the analysis so that biased estimates of safety are minimized/eliminated. It is very important to be aware of the fact that, large studies do *not* guarantee that an unbiased estimate of safety assessment will be obtained. Large studies only guarantee that the safety signal estimate will be measured with a high degree of statistical precision, ie, a small variance. Clearly, an unbiased estimate of the safety signal is what is most important to determining whether a drug should

remain or not in the marketplace. We will discuss two possible methods on how one usually attempts to obtain an estimate of safety that takes into account the issues including but not limited to patient characteristics, and selection and channeling bias. The first is regression analysis, which we will begin with. Specifically, we will discuss logistic regression since the same principles apply to all methods of regression analysis, and the dependent variable is easiest to understand since it is binary (ie, success or failure, present or absent, etc).

Logistic Regression Model

Logistic regression is one of the most frequently used statistical methods in all phases of evaluating data. It is used in almost all types of various study designs including prospectively randomized clinical trials (RCTs), observational studies, case-control and epidemiologic cohort studies. In observational studies, logistic models are frequently used to assess the contribution of risk factors to the outcome of interest. It could also be used in the controlling of imbalances between groups by adjusting for one or more covariates.[10]

A major reason to use logistic regression, is to control for confounding through consideration of many variables simultaneously. However, if there are too many variables included in the model compared with the number of cases, the estimates could be incorrect in these models[11,12]. Furthermore, the covariates are often correlated with the risk factors or predictor variables, which will lead to multiple logistic coefficients that are "collinear" and difficult to interpret.[13] Great care must be taken when using the logistic model, since it is deceptively easy to use the model in ways that are improper for its appropriate interpretation of data. These ways include issues including but not limited to: 1) not checking for model misspecification, and 2) collinearities. Another often-overlooked issue is use of the model to extrapolate to conclusions for which the data do not exist. For example, some may use the model

to make predictions for age groups not included in the actual dataset on which the analysis was based.

Thus, some have proposed use of a second method that explicitly forces the data analyst not to over-interpret the data. Also, this method attempts to make non-randomized observational/epidemiologic study data into "quasi-randomized studies" by mimicking methods found in prospectively randomized studies. This method is called the propensity score method.

Propensity Score

A relatively new statistical method is available to reduce selection and channeling bias, and to explicitly account for confounders. It also addresses the issue of inappropriate extrapolation of conclusions to places where the data do not allow valid comparisons to be made. This method is called propensity score analysis. It is increasingly used in drug safety studies, especially in large datasets. The propensity score, defined as 'the conditional probability of assignment to a particular treatment given a vector of observed covariates', was first described by Rosenbaum and Rubin in 1983.[14] It could be used in the analysis of observational and epidemiologic studies to reduce bias by identifying control subjects that are matched in probability on a large number of potential confounders to cases.[15] The propensity score will adjust the different nonrandomized groups in terms of known covariates, in order to perform between-treatment group comparisons. Thus, the propensity score is a method that allows for scientifically valid conclusions to be drawn. It is fundamentally different from regression analysis methods where such explicit matching is not performed.

There are three main approaches for taking into account confounders and/ or selection/channeling bias: matching, stratification/regression adjustment, and weighting. Matching adjusts for differences via study design, and stratification/regression adjusts during estimation of treatment effect.[16] Propensity scores represent an alternative way for adjustment for confounders and controlling for biases.

In most of the post-marketing safety surveillance situations, the data are obtained from observational or epidemiologic (ie, non-prospectively randomized) studies. Using propensity score as a covariate in the regression model for adjustment, and using it as a matching process are both very common. In these situations, propensity score is supposed to reduce the bias due to non-comparability between groups in the confounding variables. The goal is to obtain less biased/unbiased estimates of treatment effect. For the first condition, the propensity score serves as a covariate that indicates the probability of treatment that will be applied in a certain case. In the second case, the Propensity Score Matching (PSM) employs a predicted probability of group membership, to create treatment groups that are balanced on covariates even a large number of covariates (i.e., confounders). The PSM is generally considered the preferred approach to adjustment and minimizing selection/channeling and other biases.

Regulatory Issues and Propensity Scores

Propensity scores are increasingly used in post-marketing surveillance studies for analysis by the FDA. However, there are some issues underlying the use of this method, not only in the regulatory perspective, but also on the statistical side. These issues include the design and analysis of studies that appropriately collect information on as many important factors as

possible. This will allow the propensity score to be calculated and achieve a better covariate/confounder balance between treatment groups. This will enable academicians, pharmaceutical sponsors, and regulators around the world achieve the goal of accurate safety assessments of drugs. For it is the goal of regulators to keep safe drugs on the market, and to remove unsafe drugs from the market. It can happen however, that misuse of regression and propensity score methods can lead to safe drugs being wrongfully removed from the marketplace or unsafe drugs remaining on the market. This is where the pharmaceutical sponsors and the FDA's regulatory requirements align completely.

FDA's Opinions on Using Propensity Score

Generally speaking, there are various epidemiologic and statistical methods to identify and handle confounding in pharmacoepidemiologic safety studies. The FDA does not endorse or suggest any particular method. FDA encourages the continued development, use, and evaluation of innovative methods for confounding adjustment.[17]

The FDA's view on use of the propensity score issue may be summarized as follows. A propensity score for an individual is a predicted probability for treatment with a particular drug (usually the drug under study), which is conditioned on the measured covariates within the databases.[18] However, FDA also clarified a very clear rule: "when propensity score modeling is used, investigators should present diagnostics of the propensity score model to allow for an assessment of its performance and fit." The propensity score discussion has been discussed widely, and many articles provide a more in-depth discussion of this model and its appropriate application to pharmacoepidemiologic safety studies.[19,20,21]

From the FDA's viewpoint, there are ways to deal with confounders other than the propensity score although they may be less desirable and need to be considered on a case-by-case basis. One is to exclude patients who have risk factors for the safety outcome that are unrelated to drug use, or data were never collected. This strategy can be appropriate, but can also have unintended consequence of reducing the size of the population under study and also introducing bias if the reason for 'missingness' is not 'missing at random'. This reduces the power of the study to detect true safety signals. Another issue involves the extent to which results can be generalized when subjects are excluded, or if bias in introduced. If the reason for missing data is not 'missing at random', or MARS, then biased estimates of relative safety can arise. It is not surprising then, that the FDA discourages the exclusion of patients because it prevents investigators from enhancing the generalizability of the study results, compromises statistical power, and precludes the examination for effect modification by these other risk factors.[22]

In summary, FDA requires that all confounders, including time-varying confounders and effect modifiers, should be operationally defined and justified. Considering the rationality and practicality of this FDA reasoning, the use of the propensity score method in the controlling of confounders becomes a very attractive alternative to standard regression methods. This is especially true in the circumstance of FDA's clear opposition to the other commonly used methods.

Practical Issues in Study Design and Statistical Analysis

In the application of any statistical methodology, there are statistical as well as regulatory issues that arise in the study design and analysis process of the study results. For example, pre-specifying clinically relevant covariates are needed so that they will be measured and included in the analytic database. In addition, appropriate patient populations are needed to be identified, and essential elements of statistical analysis, planning sample size in the context of propensity score methodology should be further quantified.[23] Furthermore, missing covariates in generating propensity scores should be dealt with, and assessing the success of the propensity score method by evaluating treatment group overlap in terms of this methodology, will require revisiting the 'missing data' issue. [24]

Furthermore, propensity score methods can only adjust for observed covariates and not for unobserved ones. It is seriously degraded when important variables influencing treatment selection have not been collected. The ideal situation is occurs when two treatment groups overlap well in terms of the propensity scores, and all important covariates have been collected and missing data are minimal. We could then compare the two treatment groups adjusting for the PS.[25]

In addition, propensity score matching is a powerful, but imperfect surrogate for randomization. Propensity score matching can not considering factors that are unknown or not measured, while randomization tends to insure balance on both known as well as unknown covariates. With propensity score matching, not all patients can be matched as a result. So a large sample is desired to maintain the ability to insure balanced on the known factors of importance in predicting the outcome variable.

Cases of Inappropriate Use of Propensity Score within Clinical Trials that Failed to Get Approved under FDA's Regulatory Environment

1. Duke Database Propensity Score Matching Analysis

In 2013, FDA documented an Executive Summary for a first-of-a-kind transcatheter mitral valve repair system manufactured by Abbott Laboratories. This device has been reviewed by the Division of Cardiovascular Devices within the Center for Devices and Radiological Health of the Food and Drug Administration under Premarket Approval (PMA) application P100009, which was the subject of this Advisory Panel meeting.[26]

This executive summary is a representative one, and it shows clearly FDA's attitude towards the use of propensity score in an observational study. This summary not only introduced the background of several commonly used methods in medical device clinical trials, but also gave comprehensive comments on them. This signals FDA's intentions in the use of related methods in clinical trials including propensity scores.

Initially, in order to analyze the safety and effectiveness of the device, 351 patients were registered. Extensive post-hoc analyses were performed. However, in the study, FDA had examined the appropriateness and success of each step required to make the sponsor's proposed comparative analyses scientifically valid. The FDA concluded that significant problems existed at each step. From the sponsor's submitted materials, FDA thought 'with no comparator available for analysis, it is difficult to make a safety determination for use of this device in the high risk population'. For the effectiveness exploration, FDA believed that 'The Integrated High Surgical Risk Cohort has major design limitations since it was developed by pooling two individual cohorts, each with their own weaknesses, in a post hoc manner. These shortcomings

pose challenges to any consequential interpretation of data that would stand alone in support of a determination of safety and effectiveness, but do provide observations that are useful for hypothesis-generation necessary to guide future studies.'[27]

It is not surprising that FDA's negative comments led to denying approval of the device. The sponsor concluded that, because of a lack of data that could serve as a concurrent control group, it led them to look for an additional 'real world' cohort in moderate to severe MR patients who did not have surgery. As a result, the sponsor decided to use the Duke database because it contained a population with recorded MR and also it had patient outcome data. Because of the nature of clinical outcome data and the way it was obtained, the sponsor selected propensity score matching analysis as its method in order to decrease selection and channeling bias.

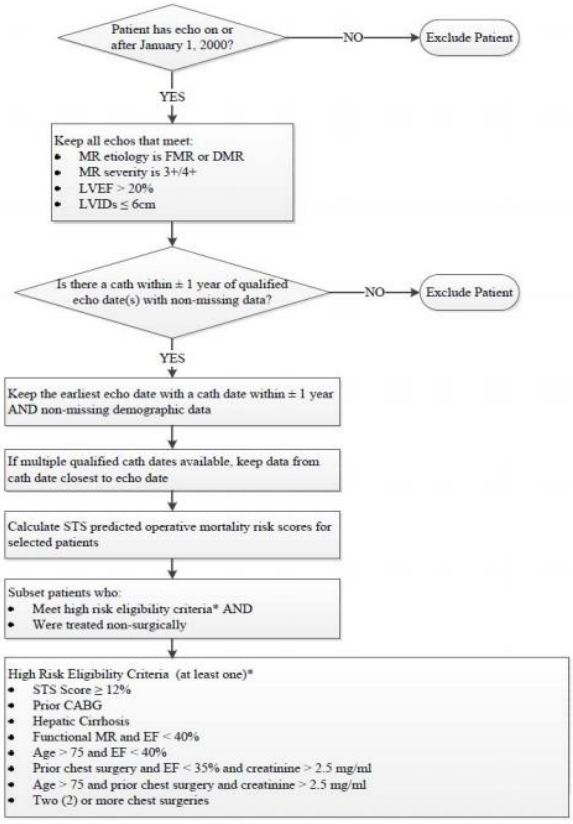


Figure 2: Duke Database Patient Selection Work Flow

It is important to emphasize that this need for re-definition of the characteristics of the Duke High Risk patients to be used for matching was identified after outcome (mortality) analysis was conducted for matched subjects. This matching allowed the patients' characteristics to more closely match the inclusion/exclusion criteria from the HRR and REALISM HR, and per the sponsor, to avoid listing of duplicate patient records.[27] This is due to the concerning of the appropriateness in the propensity score use to create balance between the two treatment groups conducted with outcome data concealed. This met the criteria and solved the potential concerns within the propensity score matching use in clinical trial studies.

The covariates listed following were specified to be included in the logistic regression in the model to calculate propensity scores:

Table 1: List of Covariates Specified to be Included in a Logistic Regression Model

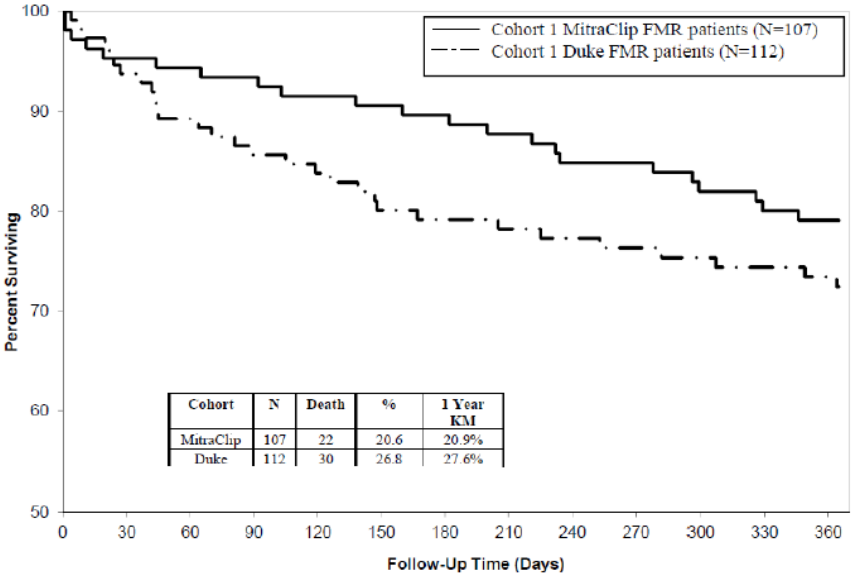
Age	Diabetes
Gender	History of renal disease
BMI	NYHA Functional Class
Previous Cardiac Surgery	History of COPD
Previous Cardiac Intervention	MR Etiology (functional or degenerative)
Previous MI	Ejection Fraction (EF)
Previous Stroke	LVIDs
History of smoking	LVIDd
History of hypertension	STS Score

From the Duke Dataset, three levels of propensity score matching were performed. However, the sponsor did not include all clinically relevant variables such as MR etiology and ventricular size in the model. As a result, functional MR can occur in non-ischemic heart failure (HF) as well and adds a high risk factor for outcomes in addition to ejection fraction (EF). The inappropriate exclusion of both MR and EF would be a serious problem even though they are closely related with functional etiology. When using propensity score for matching, both should be included.[27]

As for all the concerns, the FDA concluded that the interpretation of all the results achieved from the sponsor's analysis were neither clear nor reliable. The KM graph (Figure 3) was difficult to interpret.

Figure 3: Kaplan-Meier Freedom from All-Cause Mortality

FMR Subgroup of Matched Cohort 1



FDA's comment on this case is also negative. 'The Duke Propensity Score Analysis was retrospective, subset analysis with results that are difficult to interpret and matched cohorts do

not represent any well-defined population. Given all the concerns regarding the creation of the matched cohorts above, the difficulties in interpreting the results and considering the post-hoc nature of this analysis in a cohort that is not well-defined through different time periods, this analysis should be viewed with extreme caution and should be considered hypothesis generating. Furthermore, since the patients in the Duke databases were accumulated in an era where significant changes in medical and device therapy of CHF have occurred, there were no data on either the medical device therapy type or adequacy of treating the registry patients.

From this representative case, we can see that the FDA has a strict regulatory viewpoint and comprehensive concern about the use of propensity score in clinical studies. The population that are matched with needs to be clearly defined and reasonable, which is consistent with the concern we mentioned above.

2. A Statistical Memorandum of Alirocumab Using Propensity Score – A Case of Propensity Use in Post-Randomization and FDA's Comments

Alirocumab is Sanofi's drug whose primary treatments are hypercholesterolemia or mixed dyslipidemia. The analyses were based on data from 12 phase 2 and 3 placebo- or active-controlled trials.[28] The document illustrated the applicant's analysis methods of post-randomization subgroups based on Cox models adjusted for propensity scores.[29]

In general, propensity scores are used to conduct the trials in observational studies that lack randomization. However this one is not typical - the propensity score was used to do the post-randomization analysis.

Below are the specific questions from the FDA information request that motivated the Applicant's subgroup analyses which are evaluated.[28]

Question 2:

Provide a time to event analysis including a Kaplan-Meier cure of time to new onset of impaired fasting glucose (combining data from both AEs [adverse events] and laboratory values): (1) by treatment group and (2) within alirocumab-treated patients only, by two consecutive LDL-C values < 25mg/dL vs. others. Provide these plots for both the global pool as well as separately for the placebo and ezetimibe pools.

Question 4

Please provide in tables using the format in ISS appendix 1.4.5.4 (global pool) and 1.4.5.5 (placebo pool) TEAEs by HLGT [high level group term], HLT [high level term], and PT [preferred term] in control patients, alirocumab patients, alirocumab-treated patients with LDL-C ≥ 25 mg/dL and patients with 2 consecutive LDL-C < 25 mg/dL. Please provide p-values for the following comparisons of interest 2 LDL-C < 25mg/dL versus ≥ 25 mg/dL within alirocumab group; 2 LDL-C < 25mg/dL alirocumab versus control or placebo; and LDL-C ≥ 25 mg/dL versus control or placebo. (We recognize that this post hoc testing is exploratory and that the comparisons being made are not randomized comparisons since the subgroups are defined by post-randomization data.) Please provide a table using this same format and analyses described above listing AEs of special interest (e.g. diabetic CMQ, neurologic, neurocognitive, hepatic, etc.)

The following post-randomization subgroups were requested in the IR:

- Alirocumab LDL-C < 25 (low-LDL): patients with two consecutive LDL-C < 25 mg/dL

- Alirocumab LDL-C ≥ 25: patients without 2 two consecutive LDL-C < 25 mg/dL

For question 2, the analyses were conducted within the alirocumab arm to compare LDL-C < 25 to LDL-C ≥ 25 for patients with normal glucose at baseline or without diabetes at baseline in the specified trial groupings. For Question 4, the analyses were conducted for all patients, irrespective of baseline glucose or diabetes status, within the alirocumab arm for the trial groupings specified.[29]

However, FDA thought it was questionable that propensity scores should be used in post-randomization. As for the reviewers, they commented that the findings from analyses of post-randomization subgroups are difficult to interpret and questionable.

For example, the bias could favor the low LDL group as fewer outcomes might be considered in the analyses (e.g. outcomes that occur shortly after treatment such as injection site reactions). In addition, the bias might also disfavor the low LDL group because the follow-up period is shorter than the non-low LDL patients. Not knowing which direction the bias occurs makes it difficult to interpret if the hazard ratios obtained from the Applicant's analyses over- or under-estimate the risks for the outcomes under investigation.

Another concern is whether the propensity score analyses have adequately accounted for confounding between the alirocumab LDL-C < 25 and alirocumab LDL-C ≥ 25 groups. Typically with propensity score analyses, diagnostics are performed to assess how well the analyses have achieved its goal, i.e. to create balanced groups in terms of baseline characteristics for the comparisons. Such diagnostics have not been provided in the Applicant's response document. Therefore, there is uncertainty whether subgroup findings are due to achieving low

LDL or if due to inherent baseline characteristics of the patients that caused them to experience the outcomes analyzed.

Finally, the process for variable selection in the propensity score estimating model may not be optimal. In these analyses, prognostic factors for achieving low LDL were determined using a logistic regression model with stepwise selection for identifying factors for inclusion in the model. Stepwise selection methods have been criticized[30] for yielding inaccurate estimates of parameters and their variances. This could thereby impact the estimation of the propensity scores and lead to misclassification of patients into the quintiles used in the stratified Cox model. The consequences would include possible bias or other types of inaccuracies in the hazard ratio estimates.

The FDA concluded that given concerns with the Applicant's propensity score analyses, and concerns with analyses of post-randomization subgroups in general, there is uncertainty about the reliability of findings from these exploratory analyses. We can see from this case that the use of propensity score should be applied with great caution, such as the post-randomization analysis in this case. The safest application of propensity score analysis is in the analysis of observational studies to decrease selection bias.

3. A propensity score used in retrospective observational study in post-marketing safety surveillance

An observational study for Trasylol (asprotinin injection) was conducted by Mangano et al comprising an analysis of the international database which contained 5065 evaluable patients

collected in the multicenter study of perioperative Ischemia, Epidemiology II between 1996 and 2000.[31]

With 691 treated patients excluded, 4374 patients were categorized into one of four treatment cohorts: no treatment (1374 patients), aprotinin (1295 patients), aminocaproic acid (883 patients), and tranexamic acid (822 patients).[32]

Multivariable logistic regression and propensity-score adjustment was conducted in order to reduce bias and incorporate numerous covariates that were imbalanced between treatment groups. However, the FDA thought it was unreliable because it did not follow principles or the correct analysis of observational studies. There were several significant concerns using propensity scores used in this study:

1) The choice of covariates was apparently done explicitly using the outcome variable in a stepwise regression, thereby violating a fundamental principle of design in both randomized experiments and observational studies.[33] This will lead to exaggerated significance levels.
2) The estimated propensity score was used as a variable in a covariate adjustment and not used to create bins or to match units. This violates another rule of propensity score technology.[33]
3) It appears that distinct propensity scores were not estimated for each pair of treatment groups compared, which generally violates another rule of propensity score technology.[34]
4) The goodness of fit statistic (the C-statistic) is of limited relevance for propensity score estimation; covariate balance is critical, not fit of the underlying regression used to create the propensity score.[35,36]

The renal composite outcomes were applied in the analysis such as renal event ('renal dysfunction with an increase over preoperative baseline level of at least 62 μmol per liter' or 'renal failure requiring dialysis'), and propensity scores were used in this analysis. Except for death, the authors did not conduct any analysis to assess the risk of individual events of clinical interest.

The analysis for the renal composite outcome event in all patients of Mangano et al study [Table 2 page 359(1)], comprised comparisons between each of the three exposed cohorts (aprotinin cohort, aminocaproic acid cohort, and tranexamic acid cohort) and the untreated cohort. Odds ratios are reported after multivariable logistic regression "in the presence of covariates with propensity adjustment" based on treatment with any anti-fibrinolytic versus no treatment.

From the data, we can see that multiple significant baseline imbalances for risk factors remained for predicting renal dysfunction between the aprotinin and no treatment cohorts. So we should expect that the odds ratio for the composite renal event after 'multivariable regression and propensity adjustment' would be different from the corresponding odds ratio based on crude data. However the reported odds ratio is not very different from the odds ratio calculated from the crude data. It is not possible to make the analogous comparison for the authors' stratified analysis because the crude data Ares not available.

The propensity score in this case is not the most proper way to deal with baseline imbalances. The method it used appears to be mathematically inconsistent with the odds ratios given in the authors' Table 3[31], which reports the analysis by 'multivariable linear regression with propensity score adjustment'.

In summary, FDA commented that although Mangano et al attempted to apply propensity score methodology in the statistical analysis of this observational study, however the application was incorrect and inconsistent with its appropriate application as described by the propensity score technology as mentioned above.

Conclusions

From the cases we mentioned in this article as well as some concerns when using propensity score, we can see that FDA has strict requirements. Also considering many strict conditions for propensity score use, it should be used with great caution. When used properly, the propensity score method has great strengths in forming covariate-balanced groups between treatment groups. Then, any observed differences are likely attributable to the different treatments as compared to other covariates. We have also seen from these cases that propensity score methods have potential risks when the method is not applied correctly or in questionable circumstances. The FDA might reject a sponsor's application because of unclear pre-defined populations, or misuse to deal with imbalanced baselines. The use in post-randomization should also be careful performed.

Even though this article contains most of the concerns and some cases in the propensity scores usage under FDA regulatory environment, it did not include all the potential risks for evaluating post-marketing surveillance data. More research and reviews should be conducted to further elucidate where the use of propensity score methodology is most advantageous.

Acknowledgements

I want to thank my thesis advisors, Dr. Robert Makuch and Dr. Zuoheng Wang for their help and advices on my work. Furthermore, I also want to thank for my parents' support, not only spiritually but also their economic sponsorship. Last but not least, thank you my girlfriend and all of my friends, I cannot survive graduate school without your help and understanding.

References

[1] FDA Postmarketing Surveillance Program: http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/ucm090385.htm

[2] Brewer, Timothy, and Graham A. Colditz. "Postmarketing surveillance and adverse drug reactions: current perspectives and future needs." *Jama* 281.9 (1999): 824-829.

[3] Mathieu, Mark P., Ronald Keeney, and Christopher-Paul Milne. *New drug development: a regulatory overview*. Parexel International Corp., 2002.

[4] Brewer, Timothy, and Graham A. Colditz. "Postmarketing surveillance and adverse drug reactions: current perspectives and future needs." *Jama* 281.9 (1999): 824-829.

[5] FDA Website: Questions and Answers on FDA's Adverse Event Reporting System (FAERS): http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm

[6] Sakaeda, Toshiyuki, et al. "Data mining of the public version of the FDA Adverse Event Reporting System." *Int J Med Sci* 10.7 (2013): 796-803.

[7] Walker, Alexander M., and Meir J. Stampfer. "Observational studies of drug safety." *The Lancet* 348.9026 (1996): 489.

[8] Michels, Karin B., et al. "Prospective Study of Calcium Channel Blocker Use, Cardiovascular Disease, and Total Mortality Among Hypertensive Women The Nurses' Health Study." *Circulation* 97.16 (1998): 1540-1548.

[9] Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med 1997;127:757-763

[10] Abbott, Robert D., and Raymond J. Carroll. "Interpreting multiple logistic regression coefficients in prospective observational studies." *American journal of epidemiology* 119.5 (1984): 830-836.

[11] Harrell FEJ, Lee KL, Califf RM, et al. Regression modelling strategies for improved prognostic prediction. Stat Med 1984;3:143–52.

[12] Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable on logistic regression analysis. J Clin Epidemiol 1996;49:1373–9.

[13] Abbott, Robert D., and Raymond J. Carroll. "Interpreting multiple logistic regression coefficients in prospective observational studies." *American journal of epidemiology* 119.5 (1984): 830-836.

[14] Rosenbaum, Paul R., and Donald B. Rubin. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70.1 (1983): 41-55.

[15] Rosenbaum, Paul R., and Donald B. Rubin. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70.1 (1983): 41-55.

[16] Guo, Shenyang, and Mark W. Fraser. *Propensity Score Analysis: Statistical Methods and Applications: Statistical Methods and Applications*. Vol. 11. Sage Publications, 2014.

[17] FDA: Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm243537.pdf

[18] Rubin, D. B. (2011). "For Objective Causal Inference, Design Trumps Analysis: The Role of Propensity Scores for Design", FDA/Industry Workshop, Washington, DC

[19] Glynn, Robert J., Sebastian Schneeweiss, and Til Stürmer. "Indications for propensity scores and review of their use in pharmacoepidemiology." *Basic & clinical pharmacology & toxicology* 98.3 (2006): 253-259.

[20] D'Agostino, et al. Propensity score methods for bias reduction in the comparison of a treatment to a nonrandomized control group. Stat Med 1998;2265-2281.

[21] Esposito, et al. Results of a retrospective claims database analysis of differences in antidepressant treatment persistence associated with escitalopram and other selective serotonin reuptake inhibitors in the United States. Clin Ther 2009;31:644-656.

[22] FDA: Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm243537.pdf

[23] Yue, Lilly Q. "Regulatory considerations in the design of comparative observational studies using propensity scores." *Journal of biopharmaceutical statistics* 22.6 (2012): 1272-1279.

[24] Yue, Lilly Q. "Statistical and regulatory issues with the application of propensity score analysis to nonrandomized medical device clinical studies."*Journal of Biopharmaceutical Statistics* 17.1 (2007): 1-13.

[25] FDA Review Summary for Syncardia Systems, Inc. CardioWest Total Artificial Heart (TAH) System: Syncardia Systems, P030011. (2014)=

[26] FDA Executive Summary: Abbott Vascular MitraClip Clip Delivery System P100009. (2013) http://www.fda.gov/downloads/advisorycommittees/committeesmeetingmaterials/medicaldevices/medicaldevicesadvisorycommittee/circulatorysystemdevicespanel/ucm343842.pdf

[27] FDA Executive Summary: Abbott Vascular MitraClip Clip Delivery System P100009. (2013) http://www.fda.gov/downloads/advisorycommittees/committeesmeetingmaterials/medicaldevices/medicaldevicesadvisorycommittee/circulatorysystemdevicespanel/ucm343842.pdf

[28] FDA Statistical Memorandum: Application Number: 125559Orig1s000 http://www.accessdata.fda.gov/drugsatfda_docs/nda/2015/125559Orig1s000StatR.pdf

[29] Everett, M. B. et al. Safety Profile of Subjects Treated to Very Low Low-Density Lipoprotein Cholesterol Levels

[30] Harrell, F. E. Regression modelling strategies: With applications to linear models, logistics regression, and survival analysis. 2001. Springer-Verlag. New York.

[31] Mangano DT. Aspirin and mortality from coronary bypass surgery. N Engl J Med. 2002;347(17):1309-17.

[32] Mangano D, Tudor J, Dietzel C. The risk associated with aprotinin in cardiac surgery. N Eng J Med 2006(354):353-65.

[33] Rubin D. On principles for modeling propensity scores in medical research. Pharmacoepidemiology and Drug Safety 2004;13:855-857.

[34] Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med 1997;127(8 Pt 2):757-63.

[35] Rosenbaum P, Rubin D. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Society 1984(74):516-524.

[36] Rubin D. On principles for modeling propensity scores in medical research. Pharmacoepidemiology and Drug Safety 2004;13:855-857.