Critical Role of Statistics in Leveraging Real World Data and Evidence for Regulatory Decision-Making

Lilly Q. Yue and Heng Li

Division of Biostatistics, Center for Devices and Radiological Health,

U.S. Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993

Abstract

There are a variety of sources of real-world healthcare data that could be leveraged in the clinical studies in the regulatory settings. While such abundant data reflecting real-world clinical practice could potentially be used to reduce the cost of clinical trials, challenges arise concerning real-world data (RWD) quality, innovative analytical approaches for generating robust real-world evidence (RWE) from RWD, and appropriate use of RWE for regulatory decisions. Statistics plays a vital role in meeting all those challenges. This presentation will discuss such challenges, and the opportunities they bring about, from statistical and regulatory perspectives, illustrated with examples from medical device regulatory evaluations.

Keywords: Real-world data/evidence; Propensity score; Power Prior; Composite likelihood.

1. Introduction

In recent years, there is a growing interest in leveraging real-world data (RWD) in medical product development. Real-world data in biomedicine refer to the data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources. Examples of RWD sources include electronic health records (EHRs), insurance claims and billing data, patient registries (product or disease) and lab test databases. Here are three examples of national or international patient registries: 1) Interagency Registry for Mechanically Assisted Circulatory Support (INTERMACS), an NIH funded registry for FDA approved mechanically assisted circulatory support devices; 2) International Consortium of Orthopedic registries (ICOR); and 3) United Network for Organ Sharing (UNOS) registry. Real-world evidence (RWE) in biomedicine is the clinical evidence regarding the usage and potential benefits or risks of a medical product derived from analysis of RWD. Statistics play a critical role in the transformation from RWD to RWE to support regulatory decision-making. This paper will discuss how statistical methods are utilized to design and analyze clinical studies, when leveraging RWD. We will discuss using propensity score (PS) stratification to identify and construct a control group from RWD for a comparative investigational study. We will also discuss the application of two PS-based approaches: the PS-integrated power prior approach and the PS-integrated composite likelihood approach, to augment a single-arm investigational study with RWD, with the option of down-weighting information from the RWD. We will focus on study design.

2. Constructing a Control Group from RWD in a Comparative Study

A case study - In a prospective comparative study for pre-market approval, a left ventricular assist device was evaluated through comparison to a control group constructed by selecting patients from the INTERMACS registry who meet the inclusion-exclusion criteria of the study. The PS stratification method was then used to stratify all the patients (treated and control) into strata according to their PS such that the distribution of observed baseline covariates is similar between the treated and control patients within each stratum, leading to comparable treatment groups in terms of baseline covariates.

Formulated by Rosenbaum and Rubin (1983), the propensity score $e(\boldsymbol{X})$ for a patient with a vector $\boldsymbol{X}$ of observed baseline covariates in a comparative study is the conditional probability of receiving one treatment ($T = 1$) rather than the other ($T = 0$) given $\boldsymbol{X}$:

$$e(\boldsymbol{X}) = \Pr(T = 1 | \boldsymbol{X}).$$

Propensity score is a balancing score in the sense that conditional on the propensity score, the distribution of observed baseline covariates is the same between the treated and control patients. Therefore, among patients with the same value of propensity score, the distribution of observed covariates is the same between the two group of patients. When the propensity scores are balanced across the two treatment groups, the distribution of all the observed covariates are balanced in expectation across the two groups. In practice, the propensity score is estimated by modeling the probability of treatment group membership as a function of the observed covariates, typically via logistic regression.

The propensity score methodology refers to a collection of versatile statistical tools based on the concept of propensity score. Commonly used propensity score methods include matching and stratification on the propensity score, and inverse probability of treatment weighting using the propensity score. The methods could be used to design and analyze an observational study, mimicking some of the characteristics of a randomized controlled trial (Rubin 2001. 2007, 2008). The methods have been utilized in regulatory clinical studies for the evaluation of safety and effectiveness of medical products. In recent years, they have been used to leverage RWD in clinical studies to support regulatory decision-making.

In constructing a control group from RWD for a pre-market confirmatory investigational study, statistical and regulatory challenges can emerge. Such challenges include the potential of lower quality data in RWD, greater chance of introducing various biases in every stage and aspect of the investigational study, and the possibility of lack of objectivity in study design and thus the lack of reliability and interpretability of study results. The issues of study design objectivity could be addressed using a two-stage objective study design (Yue et al, 2014, 2016; Li et al, 2016), following Rubin's objective design principle (Rubin, 2001).

The first design stage is completed before the initiation of the investigational study. In this stage, a preliminary sample size of the investigational study is determined. It is important to identify an independent statistician at this stage to later design the study using the propensity score methodology. It is good practice for this statistician to be blinded to any outcome data, including such outcome data as have already been collected

from an RWD source like an existing patient registry, during the entire process of designing the study. Some masking mechanism, such as a firewall, is to be planned in this stage to control access of outcome data. The second design stage starts ideally as soon as patient enrollment is concluded and baseline covariate data are available for all patients. The previously identified independent statistician performs the task of estimating the propensity scores, matching patients based on the propensity scores, and assessing the covariate balance. This process is usually iterative until adequate covariate balance is reached. The selection of patients from RWD and sample size estimation are finalized at this stage along with a detailed statistical analysis plan (Yue et al 2014). During the entire study design process, only the treatment assignment and baseline covariate data are needed. Any clinical outcome data and follow-up information are neither needed nor accessed.

Illustrative Example 1. An investigational medical device was compared to a control to be selected from a national registry.

The first design stage consists of the following elements:

- Primary outcome was specified as treatment success;
- Non-inferiority margin was specified as $\delta = 11\%$;
- Propensity score stratification was planned for study design and outcome analysis;
- Independent statistician was identified;
- 15 baseline covariates were considered;
- Significance level was specified as $\alpha = 0.025$;
- Proposed sample size: N=250 for the investigational device group;
- Anticipated sample size: N=500 for the control group.

The second design stage consists of the following elements:

- Started when the enrollment of investigational study was completed;
- Based on the pre-specified patient inclusion/exclusion criteria of the investigational study, N=1,000 potential control patients were identified from the registry;
- Based on the treated patients (N = 250) and control patients (N=1,000), PS of each patient was estimated with 15 covariates included;
- PS stratification was performed.

Table 1. Distribution of patients at the five propensity score quintiles – based on 1250 patients (control: 1000; treatment: 250).

	Propensity Score Quintile					
	1	2	3	4	5	Total
Control	250	244	234	186	86	1000
Treatment	0	6	16	64	164	250

The enrollment in to the treated group stopped when 250 patients were accumulated. As it turned out, 1000 patients from the registry met the selection criteria, and were included in the study, resulting in a total sample size of 1250. A PS model was fit, and a PS stratification was done on the 1250 patients by an independent statistician who was blinded to the outcome data. Table 1 displays the number of patients in each propensity score stratum. Given that the first stratum (or PS quintile) contains no patients from the treated group, it was considered reasonable to discard the control patients in that stratum

(i.e., the first PS quintile), as they look nothing like any treated patients, with respect to propensity score and some baseline covariates. However, exclusion of any patients treated with the investigational device should be discouraged for *pre-market confirmatory studies* (we want the set of treated patients to be representative of the population of interest).

After excluding the 250 control patients in the first PS quintile, the independent statistical built a new PS model based on the remaining 1000 patients (while still blinded to the outcome data), and the number of patients in each PS quintile is displayed in Table 2.

Table 2. Distribution of patients at the five propensity score quintiles – based on 1000 patients (control: 750; treatment: 250).

	Propensity Score Quintile					
	1	2	3	4	5	Total
Control	196	193	172	128	61	750
Treatment	4	7	28	72	139	250

Treatment group comparability with respect to each baseline covariate was then assessed and thought to be satisfactory. At this time, power and Type I error rate were revisited and found to be adequate. Thus, the second design stage was completed (entirely outcome-free), and the independent statistician delivered the report of study design. The report contains all the information needed to conduct the outcome analysis including which PS stratum each patient belongs to. Each selected control patient would contribute 100% of their information.

In the outcome analysis, within-stratum comparison was made between the investigational device group and the control group based on which an overall treatment effect was estimated.

3. Augmenting a Single-Arm Investigational Study with RWD

Two PS-based methods have recently been developed for augmenting a single-arm investigational study (the current study) with RWD in the following two papers: "Propensity Score-Integrated Power Prior Approach for Incorporating Real-World Evidence in Single-Arm Studies" (Journal of Biopharmaceutical Statistics, https://doi.org/10.1080/10543406.2019.1657133) and "Propensity Score-Integrated Composite Likelihood Approach for Incorporating Real-World Evidence in Single-Arm Studies" (Journal of Biopharmaceutical Statistics, revision submitted). In those approaches, propensity score methodology is used to design a study to incorporate RWD by selecting (borrowing) comparable patients from the RWD source. The patient information borrowed from the RWD source is then down-weighted via the power prior or composite likelihood techniques to perform outcome data analysis.

According to Chen and Ibrahim (2000) a power prior is constructed as follows

$$\pi(\boldsymbol{\theta}|\, D_0, \alpha) \propto [L(\boldsymbol{\theta}|\, D_0)]^{\alpha}\, \pi_0(\theta)$$

where θ is the parameter of interest;

$L(\theta|D_0)$ is the likelihood of the external data D_0;

$\pi_0(\theta)$ is the initial prior distribution for θ;
α, the power parameter ($0 \le \alpha \le 1$), controls how much external data to borrow;
$\alpha = 0$: borrow none
$\alpha = 1$: borrow all.

Composite Likelihood (Varin et al, 2011) is a weighted product of probability density functions and takes the form:

$$L(\theta|Y) = \prod_i f(y_i|\theta)^{\lambda_i}$$

where λ_i is a nonnegative weight to be chosen to discount patient information from the RWD data source (For example if $\lambda_i = 0.6$, 60% of this patient's information is borrowed and 40% discounted). We set $\lambda_i = 1$, if patient *i* is from the investigational study (the current study); and $0 < \lambda_i \le 1$, if patient *i* is from the RWD source.

In this section, we again focus on study design, illustrated by an example below. Here patients from the current study are labeled Z = 1 and patients from the RWD source are labeled Z = 0, and propensity score is defined accordingly:

$$e(\boldsymbol{X}) = \Pr(Z = 1|\, \boldsymbol{X}).$$

Regarding discounting patient information from RWD source, a critical question to consider is how and when to determine the discount parameter α or $\boldsymbol{\lambda}$ for a prospective investigational study.

Illustrative Example 2. An investigational study (the current study) was planned to seek approval for indication expansion of an approved device. It is known that plenty of off-label use data were captured in a patient registry. Based on clinical and statistical evaluations, the registry was considered relevant to the current study with adequate reliability. Therefore, it was decided that some data be borrowed from the registry to save sample size required for the current study. We set:

- Primary endpoint: one-year adverse event;
- Parameter of interest: θ, proportion of patients who experienced adverse event(s) within a one-year period;
- Associated hypothesis testing:
  $Ho: \ \theta \ge 36\% \quad vs: \quad Ha: \ \theta < 36\%$;
- Study success criterion:
  - Posterior probability of θ being less than 0.36 is greater than 0.95; or
  - *p*-value < 0.05 in frequentist setting;
- 17 baseline covariates were identified based on prior clinical knowledge;
- Sample size determination
  - Assume $\theta = 0.30$;
  - Set: power = 80%; significance level = 0.05;
  - Then, N = 380;
  - Based on clinical decision (case-by-case basis), proposed to
    - *Enroll* 290 patients in the current investigational study
    - *Borrow* 90 (about 25%) patients from the registry.

After the enrollment of all 290 patients was completed in the current study, propensity score was estimated for each patient, and 941 external patients were selected using PS, by excluding those external patients whose PS is not in the range of that of the patients in the current study. All patients (290 + 941) were grouped into 5 PS strata, with the same number of current study patients (290/5 = 58) in each PS stratum. The purpose of the PS stratification is to create strata of patients such that within each stratum external patients and patients from the current study are more similar in terms of observed covariates than they are overall. Borrowing of external patients is then carried out within each stratum to make it more justified. The number of registry patients and current study patients are displayed in Table 3.

Table 3. Sample Size in PS Stratum

	1	2	3	4	5	Total
Current Study (n)	58	58	58	58	58	290
Registry (n)	281	210	154	187	109	941

Note that it was decided based on clinical considerations that only 90 external patients were to be borrowed, but 941 external patients were selected. Therefore, only a fraction of information from each of the 941 external patients can be borrowed. We consider two approaches to incorporating partial information from external patients, a Bayesian approach via power prior and a frequentist approach via composite likelihood. The fraction of information to be borrowed is controlled by the power parameter α in the former approach and the weight λ in the latter approach. Now it may seem that all we need to do is to set the value of α or λ so that the fraction of information barrowed is equivalent to 90 patients. But it is not that simple. Since borrowing takes place within PS strata, we need to figure out how to allocate the total information to borrow, which is equivalent to 90 patients, to each PS stratum.

There are many possible ways to allocate the nominal number of 90 patients into 5 PS strata. Our strategy is to make the nominal number of patients to be borrowed in each stratum *proportional* to the similarity of RWD patients and the current study patients in terms of baseline covariates in that stratum, where similarity is measured by an *overlapping coefficient,* the overlapping area of propensity score distributions of the two groups of patients. The overlapping coefficient in each stratum is displayed in Table 4. The overlapping coefficients are then standardized so that they add up to 1. The standardized overlapping coefficients times the total number of patients to be borrowed (90) determine the nominal number of patients to be borrowed in each stratum (E.g., in the first PS stratum, 90 x 21% = 19).

Table 4. Determination of Power Parameter or Weight in Each PS Stratum.

	1	2	3	4	5	Total
Current Study (n)	58	58	58	58	58	290
Registry (n)	281	210	154	187	109	941
Overlap Coeff.	0.87	0.78	0.86	0.84	0.77	
Std. Overlap Coef.	21%	19%	21%	20%	19%	100%
Patients Borrowed	19	17	19	18	17	90
α (or λ)	0.07	0.08	0.12	0.10	0.15	

In each PS stratum, the power parameter α in the Bayesian approach, or the weight λ in the frequentist approach, can then be obtained by dividing the nominal number of external patients to be borrowed by the total number of RWD patients in that stratum (E.g., 19/281 = 0.07 in the first PS stratum). Having determined α (or λ) in each PS stratum, we know the fraction of information each external patient contributes, and the study design is complete. Here, again, all the above design activities were performed by an independent statistician who was blinded to the outcome data.

After clinical outcomes were observed from all the patients, the final analysis was conducted, based on the PS study design. For the Bayesian approach, apply the power prior within each stratum to get stratum-specific posterior distributions, which are then combined to complete the inference for the parameter of interest. In this example, the posterior probability of $\theta < 36\%$ is 96.9%, which meets the study success criterion. For the frequentist approach, construct the composite likelihood to get stratum-specific parameter estimates, which are then combined to complete the inference for the parameter of interest. In this example the maximum likelihood estimate of $\theta = 31\%$, p - value = 0.01.

4. Concluding Remarks

High quality real-world data have the potential to play an important role in regulatory decision-making. In this paper we discussed statistical approaches that can be used to incorporate RWD in prospective clinical studies, demonstrated by illustrative examples. In the first example patients from a registry are used to constitute the control group of a comparative device study, with the investigational device arm consisting of prospectively enrolled patients. The second example is a single-arm study in which RWD from off-label use are synthesized with prospectively enrolled patients to form a single arm study to support indication expansion for an approved device. In both cases the study design is based on propensity score methodology, which is utilized to balanced baseline covariates between the prospectively enrolled patients and those from an RWD source. In the comparative study the objective of covariate balancing is for making causal inference for the treatment effect of an investigational device versus standard of care (the control). In the single arm study covariate balancing is used for making the leveraging of real-world data more justified. Besides using propensity score (stratification) to balance covariates, the study design in the second case also involves the determination of the total nominal number of external patients to borrow, the nominal number of patients to borrow in each propensity score stratum, and the weights used to down-weight information contributed by external patients. Note that all the design activities can and should be carried out without any outcome data in sight. It is important to ensure that the study design is outcome-free to safeguard objectivity and integrity so that study results are credible. The implementation of such outcome-free design depends on the cooperation of multiple stakeholders with the statistician playing a central role (Xu et al, 2019). More generally, statisticians can make myriad contributions in the transformation of RWD into RWE. Our case studies are just two examples of such contributions. The methods we have developed can readily be used in practice, and we also hope that they can serve as stepping stones for more research in the future.

Acknowledgements

We'd like to acknowledge our RWD/RWE research collaborators in JHU, Dr. Chenguang Wang, and, at the FDA, Drs. Wei-Chen Chen, Nelson Lu, Ram Tiwari and Yunling Xu.

References

1. Austin, P. (2011). An introduction to propensity score methods for reducing the effect of confounding in observational studies, *Multivariate Behavioral Research,* 46:399-424.
2. Chen, M-H and Ibrahim, J.G., (2000) Power Prior Distribution for Regression Models. Statistical Science, 15(1): 46-60.
3. Li, H., Mukhi, V., Lu, N., Xu, Y. & Yue, Q.L. (2016). A note on good practice of objective propensity score design for premarket nonrandomized medical device studies with an example. Statistics in Biopharmaceutical Research, 8(3), 282-286,
4. Rosenbaum, PR, Rubin DB (1984). Reducing bias in observational studies using subclassification on the propensity score. *JASA*, 79:516-524.
5. Rubin, D.B. (1997). Estimating casual effects from large data sets using propensity scores. *Ann Intern Med*, 127:757-763.
6. Rubin, D.B. (2001). Using propensity score to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology,* **2**, 169-188.
7. Rubin, D.B. (2007). The design versus the analysis of observational studies for causal effects: Parallel with the design of randomized trials. *Statistics in medicine*, **26**: 20-36.
8. Rubin, D.B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2 (3), 808-840.
9. Varin et al (2011). An overview of composite likelihood methods. Statistics Sinica, 5-42.
10. Wang, C., Li, H., Chen, W., Lu, N., Tiwari, R., Xu, Y., Yue, L. (2019). Propensity Score-Integrated Power Prior Approach for Incorporating Real-World Evidence in Single-Arm Clinical Studies. *Journal of Biopharmaceutical Statistics, https://doi.org/10.1080/10543406.2019.1657133*.
11. Wang, C., Lu, N., Chen, W., Li, H., Tiwari, R., Xu, Y., Yue, L. (2019). Propensity Score-Integrated Composite Likelihood Approach for Incorporating Real-World Evidence in Single-Arm Clinical Studies. *Journal of Biopharmaceutical Statistics,* revision submitted.
12. Xu, Y., Lu, N., Yue, L.Q., Tiwari, R. (2019). A Study Design for Augmenting the Control Group in a Randomized Controlled Trial: A Quality Process for Interaction Among Stakeholders. *Therapeutic Innovation and Regulatory Science*, 1-6.
13. Yue, L.Q., Lu, N. and Xu, Y. (2014). Designing pre-market observational comparative studies using existing data as controls: challenges and opportunities. *Journal of Biopharmaceutical Statistics* 24: 994-1010.
14. Yue, L.Q., Campbell, G., Lu, N., Xu, Y., & Zuckerman, B. (2016) Utilizing national and international registries to enhance pre-market medical device regulatory evaluation. *Journal of Biopharmaceutical Statistics,* 26 (6), 1136–1145.