Multinational Real-world Studies in the Biopharmaceutical Industry: Design, Analysis, Issues, and Case studies

Macaulay Okwuokenye
Syros Pharmaceuticals Inc.;
Adjuct Faculty Biostatistics, JPHCOPH,
Georgia Southern University and University of New England

BASS XXV
October 16, 2018

Acknowledgement

Thanks for your audience

Disclosure

Contents herein are solely the responsibility of the author and does not represent the views of author's employers

Outline

1. Outline
2. Real-world data and Real-world Evidence
3. Multinational Real-world Studies
4. Designing Real-World Studies
5. Propensity Score
    - Propensity Score Methods
6. Case Study I:
7. Case Study II: Optimal Time to Switch Therapy
8. Biomarker Screening in Early Phases
9. Sensitivity Analysis
10. Publishing Results

Outline

Outline:

- Real-world data and real-world evidence
- Role of real-world data in evidentiary inference
- Multi-national real-world studies
- Design and analysis of credible real-world clinical studies
- Methodological aspects of study design and analysis
    - Propensity score methods
    - Case Studies
- Issues
- Sensitivity analysis
- Publishing results

Table of Contents

1. Outline
2. Real-world data and Real-world Evidence
3. Multinational Real-world Studies
4. Designing Real-World Studies
5. Propensity Score
   - Propensity Score Methods
6. Case Study I:
7. Case Study II: Optimal Time to Switch Therapy
8. Biomarker Screening in Early Phases
9. Sensitivity Analysis
10. Publishing Results

Real-world data and Real-world Evidence

Real-world data (RWD) are data obtained from sources that are departures from traditional randomized controlled double blinded experiment (RCDBE).

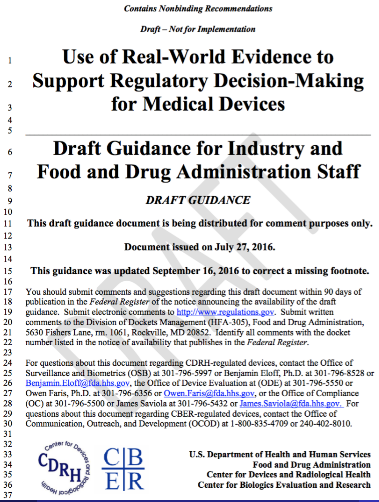
- Departures could include non-randomized, non-controlled, non-double blinded, or any combination of these.
- Example data sources:
    - Large simple trials or pragmatic trials
    - Retrospective observational or registry studies
    - Case reports
    - Administrative and healthcare claims and electronic medical records
    - Some genomic studies
    - Other observational study settings
- Real-World Evidence (RWE) is any evidence gathered from RWD.
- RWE can potentially complement information from traditional randomized clinical trials

Real-world data and Real-world Evidence

- Depending on data characteristics, RWD could be a valid source of scientific evidence (FDA Draft guideline for RWD [1]).
- Real-world studies–if well designed, conducted and analyzed–could provide insights on causal treatment effects in settings where feasibility and ethics preclude conduct of randomized clinical trials. Examples:
    - Smoking and lung cancer
    - Auto accident fatality and use of car seat beat
    - Optimal time to switch patient to a new therapy when treatment holiday might be unethical
    - Comparative effectiveness of therapies in routine clinical practice post-approval of a drug
    - Comparison of new medical device to existing one

Real-world data and Real-world Evidence

- In July 2016, the FDA issued a draft guideline on use of real-word evidence to support regulatory decision for medical devices



Contains Nonbinding Recommendations

Draft – Not for Implementation

Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices

---

Draft Guidance for Industry and Food and Drug Administration Staff

DRAFT GUIDANCE

This draft guidance document is being distributed for comment purposes only.

Document issued on July 27, 2016.

This guidance was updated September 16, 2016 to correct a missing footnote.

You should submit comments and suggestions regarding this draft document within 90 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to http://www.regulations.gov. Submit written comments to the Division of Dockets Management (HFA-305), Food and Drug Administration, 5630 Fishers Lane, rm. 1061, Rockville, MD 20852. Identify all comments with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions about this document regarding CDRH-regulated devices, contact the Office of Surveillance and Biometrics (OSB) at 301-796-5997 or Benjamin Eloff, Ph.D. at 301-796-8528 or Benjamin.Eloff@fda.hhs.gov, the Office of Device Evaluation at (ODE) at 301-796-5550 or Owen Faris, Ph.D. at 301-796-6356 or Owen.Faris@fda.hhs.gov, or the Office of Compliance (OC) at 301-796-5500 or James Saviola at 301-796-5432 or James.Saviola@fda.hhs.gov. For questions about this document regarding CBER-regulated devices, contact the Office of Communication, Outreach, and Development (OCOD) at 1-800-835-4709 or 240-402-8010.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Devices and Radiological Health
Center for Biologics Evaluation and Research

Table of Contents

1. Outline
2. Real-world data and Real-world Evidence
3. Multinational Real-world Studies
4. Designing Real-World Studies
5. Propensity Score
    - Propensity Score Methods
6. Case Study I:
7. Case Study II: Optimal Time to Switch Therapy
8. Biomarker Screening in Early Phases
9. Sensitivity Analysis
10. Publishing Results

Multi-national Real-world Studies

Multi-national real-world studies are real-world studies conducted across multiple countries:

Attractive because they:

- potentially increase access to eligible patients/participants
- possibly quicker study completion timeline; therefore, faster data to decision
- potentially improved generalizability of results since patients are recruited from different countries and ethnicities.
- are not without issues

Table of Contents

1. Outline
2. Real-world data and Real-world Evidence
3. Multinational Real-world Studies
4. Designing Real-World Studies
5. Propensity Score
    - Propensity Score Methods
6. Case Study I:
7. Case Study II: Optimal Time to Switch Therapy
8. Biomarker Screening in Early Phases
9. Sensitivity Analysis
10. Publishing Results

Designing Real-World Studies

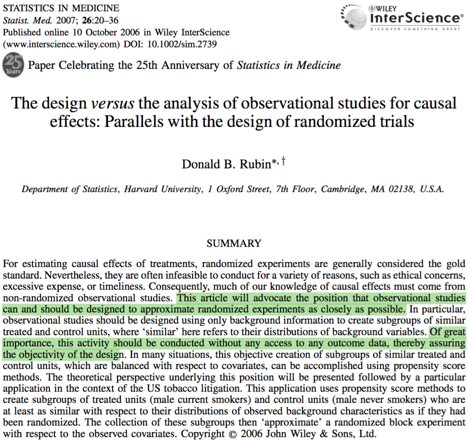
- Consider a study to compare outcome of an intervention between two treatment groups?
  - A fair comparison of causal association on an outcome between treatment groups will ensure groups similarity pre-treatment

Randomization ensures similarity (on average) between groups in terms of measured and unmeasured pre-treatment variables

- Randomization justifies valid inferential comparison between treatment groups without recourse to complex mathematics
  - Assuming study is not compromised by post-randomization events, e.g., differential loss to follow-up
- Non-randomized studies lack the benefits of well designed and conducted randomized experiments:
  - A naive comparison of unbalanced treatment groups will be invalid and uninterpretable
  - Association may be due, among many things, to: bias, chance, confounding, and cause

Designing Real-world Studies

- - Good design and analysis aim to prevent, reduce, and evaluate bias, confounding, and chance, to enable estimation of a causal unbiased association between exposure and outcome [2].



STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:20–36
Published online 10 October 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2739

Paper Celebrating the 25th Anniversary of *Statistics in Medicine*

The design *versus* the analysis of observational studies for causal effects: Parallels with the design of randomized trials

Donald B. Rubin*,†

Department of Statistics, Harvard University, 1 Oxford Street, 7th Floor, Cambridge, MA 02138, U.S.A.

SUMMARY

For estimating causal effects of treatments, randomized experiments are generally considered the gold standard. Nevertheless, they are often infeasible to conduct for a variety of reasons, such as ethical concerns, excessive expense, or timeliness. Consequently, much of our knowledge of causal effects must come from non-randomized observational studies. This article will advocate the position that observational studies can and should be designed to approximate randomized experiments as closely as possible. In particular, observational studies should be designed using only background information to create subgroups of similar treated and control units, where 'similar' here refers to their distributions of background variables. Of great importance, this activity should be conducted without any access to any outcome data, thereby assuring the objectivity of the design. In many situations, this objective creation of subgroups of similar treated and control units, which are balanced with respect to covariates, can be accomplished using propensity score methods. The theoretical perspective underlying this position will be presented followed by a particular application in the context of the US tobacco litigation. This application uses propensity score methods to create subgroups of treated units (male current smokers) and control units (male never smokers) who are at least as similar with respect to their distributions of observed background characteristics as if they had been randomized. The collection of these subgroups then 'approximate' a randomized block experiment with respect to the observed covariates. Copyright © 2006 John Wiley & Sons, Ltd.

Designing Real-World Clinical Studies

- Real-world clinical studies should be designed to closely mimic simple randomized trial:
  - Protocol: developed before study begins
    - Specify the causal association/treatment effect of interest
  - SAP: written and signed off before data base lock/data transfer.
    - Careful thought warranted before specifying any analysis in the SAP; analyses could become burdensome.
  - Study design is separated from and precede data analysis.
  - Outcomes are withheld from analyst during the study design.
  - Treatment and time of treatment initiation are clearly defined.
  - Variables measured before and after treatment initiation are clearly distinguished.
  - Variables measured after treatment initiation are never used in study design.

Designing Real-World Clinical Studies

- Decision level that impacts treatment assignment are critical–e.g., types of insurance, type of healthcare, first-line versus second-line therapy, prior therapy, etc.
- Reduce sample heterogeneity using inclusion/exclusion criteria; nonetheless, consider impact on generalizability.
- Ensure treatment groups are balanced on measured covariates; however, this provides no basics for anticipating balance in unmeasured covariates.
- Inability to induce balance among unmeasured covariates is an important limitation of non-randomized studies.
    - Among methods of treatment assignment, randomization boast of ability to induce balance among unmeasured covariates.

Designing Real-World Clinical Studies

- Collect data on possible alternative explanation for treatment effects:
  - e.g., using multiple control groups that are subject to different source of biases,
  - or baseline pre-treatment measurements of the outcome,
  - or data from different region (one can assess effect of same treatment but from different region on the outcome).
- Inclusion/exclusion criteria are based on variables measured before treatment initiation.
- Plan for sensitivity analysis.
- Identity a control or reference group (See [3] for guideline).
  - In selecting control group, consider possible impact of temporal effects.
  - This is important in instances that experienced rapid technological evolution or in diseases where natural remission is possible.

Control Group

- A reference or control group is important.
    - It provides a comparison group.
- There are various types of control.
    - Concurrent controls (e.g., prospective studies involving two arms; placebo).
    - Self-control (comparison with pre-treatment/baseline value; pre-post).
        - Interpretation should consider the critical role of possible regression to the mean, particularly in remitting diseases.
        - One could use multiple periods, e.g., before treatment, during washout, during treatment, and after treatment.

Designing Real-World Clinical Studies

- Historical control
  - One might use patient-level data from prior trials.
  - When patient level data are unavailable, a fixed target may serve as reference for comparison.
  - Meta-analyze multiple historical trials data to obtain a distribution of effects.
  - Be careful when there is possible natural temporal change in disease course or temporal evolution of technology (e.g., if disease rate/incidence is known to have changed over time, or diagnostics improvement).

Designing Real-World Clinical Studies: Propensity Score

The Propensity Score

- Denote by $\boldsymbol{X}$ a covariate vector for a subject assigned to treatment T (=1 if treated; 0 if untreated). The propensity score [4], $e(\mathbf{x})$, is the conditional probability of treatment assignment given observed covariates; that is,

$$e(\mathbf{x}) = Pr\left(T = 1|\boldsymbol{X}\right); \tag{1}$$

Designing Real-World Clinical Studies: Propensity Score

- Equation (1) implies that treated and untreated subjects with the same value of propensity score will have same distribution of observed covariate $\boldsymbol{X}$.
- This is formally expressed as the treatment assignment and the covariates been conditionally independent given the propensity score.
- Propensity score is a balancing score.
  - Propensity score is used to induce balance between intervention groups on observed pre-treatment covariates.

Designing Real-World Clinical Studies: Propensity Score

- Propensity score methods are appropriate when randomization into groups to be compared could have been feasible:
    - e.g., intervention group, and any categorization defined on function of time before commencing treatment–early versus later starter.
    - i.e., when manipulation of treatment by an investigator would have been possible.
- The true propensity score is rarely known except in randomized experiments:
    - It has to be estimated estimated in other settings.

Estimating Propensity Score

- The goal is to estimate propensity score that balance covariate distribution.
- The goal is NOT to obtain the best estimate of propensity score in terms of any criteria based on minimizing the difference between estimated and true propensity score.
- Some methods of using propensity score more than others may rely on an accurate approximation of the true propensity score by the estimated propensity score.
    - Hence, decision made during propensity score model specification may impact estimators for treatment effects.
    - Justification for the final propensity score model should be clearly stated.

Using The Estimated Propensity Score

Propensity score Methods: Stratification

- Stratification involves using the estimated propensity score to partition/separate subjects into mutually exclusive strata.
- Within each strata with similar values of propensity score, the treatment and control subjects have similar covariate distribution.
- Stratification with further within-strata model-based covariate adjustment will less likely require exact specification of propensity score model.
- The number of strata is data-driven; quintile is common.
- Overall treatment effects could be obtained as a weighted average of stratum-specific estimates.

Table of Contents

Case Study I: Comparative Effectiveness Post-approval

- Study Objective:
  - To evaluate 12 months on-treatment comparative effectiveness of two therapies in routine clinical practice post-approval
- Population:
  - Defined through inclusion/exclusion criteria to reflect the population of interest including population excluded from the pivotal trials
- Post-treatment initiation event
  - None considered during study design stage
    - Study duration was restricted to 12 month to minimize complication created by post-treatment initiation events
    - Treatment discontinuation was not expected to be much given what was known about the two therapies
    - Patients were censored at time of switch to alternative therapy

Study Design Considerations

- Variables
  - Time to relapse
  - Number of relapse
- Primary endpoint
  - Proportion of patient who relapsed
- Secondary endpoint
  - Frequency of relapse
- Summary measure
  - Hazard ratio
  - Kaplan-Meier estimate of proportion relapsed
  - Annualized relapse rate at 12 months

Study Design

- PS estimated with baseline variables based on substantive ground
  - age, sex, region, number of relapses in the past year, time since first MS symptoms, number of prior MS treatments, reason for discontinuing prior MS treatment, missing baseline EDSS score indicator.
  - Region was grouped based on similarity in heathcare system
- Remaining pre-treatment variables were sequentially added to the PS model at P-value=0.2 threshold
- Quartiles (strata) of PS were created

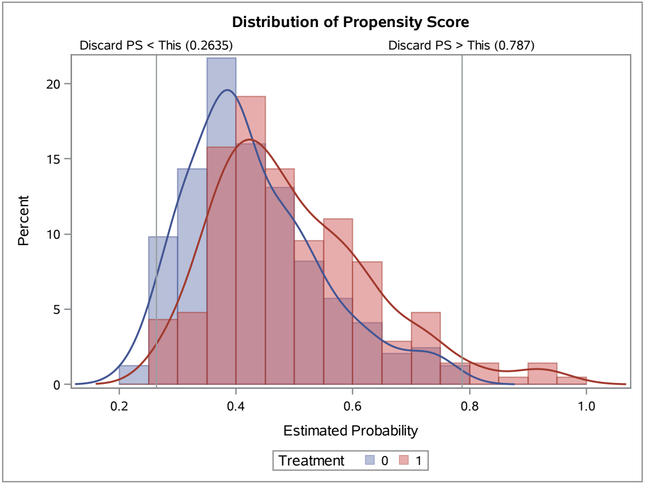
Stratification

Assessing Adequacy of Number of strata:

- The goal was to construct subclasses within which there was a modest variation in estimated propensity score.
  - Within each strata, the treatment indicator should be statistically approximately unrelated to the estimated propensity score.
- Estimated propensity score were trimmed to avoid ”prophetic extrapolation” (i.e., induce overlap in propensity score distribution):
  - Discarded control units having an estimated propensity score less than the smallest value of estimated propensity score among treated subjects
  - Discarded treated units having an estimated propensity score greater than the estimated propensity score in control subjects

Trimming PS

1



Stratification

- Compute estimated linearized propensity score $\hat{\ell}(x)$

$$\hat{\ell}(x) = \ln\left(\frac{\hat{e}(x)}{1-\hat{e}(x)}\right) \tag{2}$$

- Assess whether linearized propensity score $(\hat{\ell}(x))$ is approximately unassociated with treatment indicator within strata
  - t-test were used to assess the hypotheses that $\hat{\ell}_{tj}$ and $\hat{\ell}_{cj}$ are from same distribution
  - $$t - stat_j = \frac{\bar{\ell_{tj}} - \bar{\ell_{cj}}}{\sqrt{S^2_{\ell j} \times (1/N_{cj} + 1/N_{tj})}} \tag{3}$$
  - where in stratum j, $\bar{\ell_{tj}}$ and $\bar{\ell_{cj}}$ are the means of the linearized propensity score $(\hat{\ell}(x))$; S^2_{lj} is the sampling variance of $\hat{\ell}(x)$, and N_{tj} and N_{cj} are the number of treated and control subjects, respectively.
  - If t-statistics is larger than a specified threshold (e.g., $t_{max} = 1$) and the sample size is sufficiently large, then split the strata by median propensity score in that strata.

Evaluating Adequacy of Propensity Score Model

- Displayed distribution of propensity score by strata–histogram and box plot.
- Normalized difference (Ndif) was computed for continuous covariates; a Ndif of 0.1 was deemed a satisfactory balance.
- A 2-way ANOVA with the covariate as a pseudo-outcome was used to assess the balance in baseline continuous covariates; a logistic or Poisson regression for dichotomous and count variables, respectively.

Evaluating Adequacy of Propensity Score Model

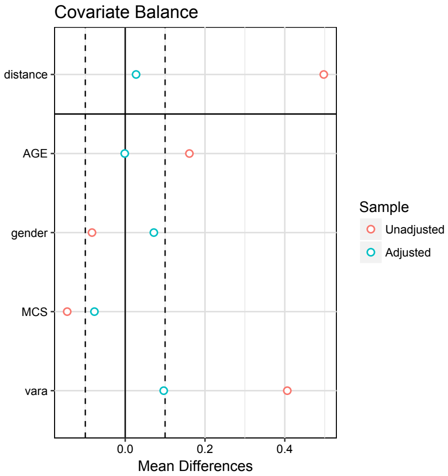


Figure 2: An Example Plot of Normalize Difference

Statistical Analysis of Outcome

Upon satisfactory balance in distribution of baseline variables:

- Within each strata of PS, the proportion of patients relapsed was estimated using the Kaplan-Meier product limit method, based on a time-to-first relapse survival distribution.
- Pooled Kaplan-Meier estimates were obtained as weighted KM estimates over strata with the weight proportionate to the number of patients at risk for each interval within each stratum
  - This is essentially meta-analysis of the Kaplan-Meier estimates across strata of PS
- Hazard ratio was estimated using stratified Cox's proportional hazard model with the PS quartile as stratifying variable, adjusting for age, region, and no. of relapse in past one year
  - Further adjustment was to account for possible residual imbalance
- Analysis for annualized relapse rate was similar

Propensity Score Methods: Weighting

Propensity Score Weighting

Inverse-probability-of-treatment weight (IPTW) estimator (for estimating ATE) and ATT-weighted estimators.

- IPTW estimates standardized effects measure with the entire study population as the standard population.
- Assuming no unmeasured confounder, IPTW estimates the causal treatment effect in a population whose distribution of risk factors is same as that of entire study population.
- IPTW uses reciprocal of estimated propensity score for treated patients ($1/\hat{e}(\boldsymbol{X})$) and inverse of 1 minus estimated propensity score ($1/(1-\hat{e}(\boldsymbol{X}))$) for untreated patients.

Propensity Score Methods: Weighting

- ATT-weighted analyses uses the number 1 for the treated and propensity odds $(\hat{e}(\boldsymbol{X})/(1-\hat{e}(\boldsymbol{X})))$ for the untreated.
- ATT-weighted approach estimates a standardized treatment effects measure in which the treated group is the standard population.
- No weighting technique is superior to the other; each technique answer different questions.
- This emphasizes the importance of clearly defining the estimate of interest and justification.
- Extreme weight can impact analysis results.
- Trim extreme weights

Propensity Score Methods: Weighting

Aside:

- With successful matching of the treated subjects to the control subjects, estimates of causal association should be similar to that from ATT-weighted estimator; however, the advantages of using weighted approach are:
    - 1) It uses data from all patients
    - 2) It is unaffected by uncontrolled confounding due to inability to find a good match for treated subjects
- Weighting can also be viewed as the limit of stratification as the number of observations and subclasses tend to infinity [7]

Table of Contents

1. Outline
2. Real-world data and Real-world Evidence
3. Multinational Real-world Studies
4. Designing Real-World Studies
5. Propensity Score
   - Propensity Score Methods
6. Case Study I:
7. Case Study II: Optimal Time to Switch Therapy
8. Biomarker Screening in Early Phases
9. Sensitivity Analysis
10. Publishing Results

Case Study II: Optimal Time to Switch Therapy

Call these drug TC and NH. Following TC approval, clinicians were interested in:

- optimal time to transition patients from NH to TC;
- which patients are good candidate to transition;
- timely information on transitioning guideline

Moreover,

- lots of the patients were considerably older (as old as 70 years) with commorbidities (hence, excluded in the pivotal trials)
- Reasons for therapy switch varies among patients–patient's or investigator's decision(This is CRITICAL, we will come this!)

Case Study II: Design Consideration

Where we to conduct a CRT, a design option will be to randomize patients into early versus late transition such that $0 < Pr(T = 1|X) < 1$ hold

- Investigator would randomly assigned patients with a known and equal probability to early versus late transitioners
- Random treatment assignment justifies many of the statistical analyses approach

However,

- patient transitioned for different reasons; hence, the probability of transitioning (assignment) to early versus late is unknown
  - This probability might be confound by unobserved covariate
  - This is the reason removing bias and confounding are very important (Hint: No systematic error in the sample size formula?)
- if a driver's destination is New York; driving the car toward Maine will not get to the destination no matter the speed

Study Objective and Endpoint

- Objective
  - Compare frequency of relapse at one year post-treatment initiation among early versus late switchers, defined as ≤ 90 days versus > 90 days washout duration
  - based on the lower limit suggested for a 12-to-16 week NH washout
- Endpoint
  - Number of relapse at one year post-treatment initiation
  - Proportion of patients relapse
- Summary measure
  - Annualized relapse rate
  - Hazard ratio
  - Kaplan-Meier estimates of proportion relapse

Study Design

- More of less of case series in which each patients served as his/her control
- Phase 4 retrospective observational study of patients with RRMS who switched from NH to TC.
- Data were collected before NH initiation to obtain an estimate of ARR before and during the 1 year after NH initiation, and in the year after TC initiation
- Patients must had received $\geq$ 12 months of NH and no other TC between NH discontinuation and TC initiation
- Patient must have initiated TC $\geq$ year before study enrollment, but not required to be on treatment for the entire one year
    -

Study Design Continued...

- Baseline characteristics (age, number of relapses during NH treatment and washout, duration of NH treatment, and steroid use during washout) of these 2 groups were balanced using PS.
- A stabilized inverse probability of treatment weight (SIPTW) estimator was used to assess ARR in these 2 groups.
- The IPTW estimates a standardized effect using the entire study population as the standard population (Sturmer et al., 2010).
- Balance was assessed using Kolmogorov-Smirnov test for continuous variables and the chi-square test for categorical variables.
- A satisfactory balance was achieved in the baseline characteristics for the washout duration categories

Statistical Analysis of Outcome

- The estimated hazard ratio (HR) for impact of washout duration ($\leq$ 90 vs 90 days) among patients without relapse during natalizumab treatment was based on a Cox proportional hazard model using the SIPTW estimator.
- Variability of the HR was based on a robust (sandwich) standard error.
- Supplementary analysis (post-hoc): ARR by quartile of washout duration
- A standardized effect using the population with washout duration $\leq$ 90 days as the standard population (ATT) was also estimated (the result was similar to that based on IPTW above.

Multi-national Studies: Issues

- Data quality and reliability
  - Data quality issues often preclude conducting all the necessary analyses–sensitivity analysis
- Possible unstandardized adjudication and assessment of outcome across countries, e.g., adjudication of relapse is unlikely to be same across clinics in different countries
  - What then is the interpretation of the outcome?
- Frequency of clinic visits might not be same across countries–this could potentially affect the reported outcome
- Standard of care might differ across countries; this impact treatment assignment and prescription behavior–possibly inducing systematic differences in patient population
- Types of healthcare and insurance may differ across countries, with possible impact on prescription behavior–a possible source of systematic difference.

Multi-national Studies: Issues Continued...

- Expected outcomes for standard of care may change over time at different rate across countries, due, for e.g., to change in supportive care, differences in radiological assessment techniques, difference in availability of second line therapy
    - Population been compare might be different: e.g., patients on first-line therapy are generally different from patients on second-line therapy.
- Possible inter-institution and inter-country variability in outcomes
    - Explained variability in outcomes by measured variables may be limited; this may limit the applicability of any covariate adjustment method.

Multi-national Studies: Issues Continued...

- Validity of study design–data availability (blinding is difficult perform and document)
    - Providing evidence that outcome was not used during study design is difficult
- Missing data
    - Define what comprised missing data carefully
    - Consider data imputation carefully
    - Sometimes the estimand of interest is unclear

Table of Contents

1. Outline
2. Real-world data and Real-world Evidence
3. Multinational Real-world Studies
4. Designing Real-World Studies
5. Propensity Score
    - Propensity Score Methods
6. Case Study I:
7. Case Study II: Optimal Time to Switch Therapy
8. Biomarker Screening in Early Phases
9. Sensitivity Analysis
10. Publishing Results

Biomarker Screening in Early Phases

- Many early phase biomarker single arm studies are not randomized
  - A study might identify differential gene expression or specific mutation as prognostic or predictive
  - Patients typically are not assigned to a gene or mutation
- Unobserved covariates (e.g., age, prior therapy) might impact differential gene expression or mutation status
- Inferential approaches for real-world studies could (and should) be useful in early phase biomarker studies

Table of Contents

1. Outline
2. Real-world data and Real-world Evidence
3. Multinational Real-world Studies
4. Designing Real-World Studies
5. Propensity Score
   - Propensity Score Methods
6. Case Study I:
7. Case Study II: Optimal Time to Switch Therapy
8. Biomarker Screening in Early Phases
9. Sensitivity Analysis
10. Publishing Results

Sensitivity Analysis

- Sensitivity analysis involves assessing the sensitivity of conclusions to assumption about unmeasured covariate.
- One of the earliest is due to Cornfield et al (1959) [8]
  - They considered sensitivity of conclusion to unmeasured binary covariate
- Method due to Rosenbaum (1995) sensitivity is common:
  - Rosenbaum calculated Fisher p-value for assessing fisher sharp null hypothesis of no treatment effects
  - Assess the sensitivity of the conclusion under unconfoundness to that assumption
- The method due to Lin, Psaty, and Kronmal (1998) could be applied to model-based treatment effects estimates.
  - This could easily be used to assess sensitivity for summary published results

Table of Contents

1. Outline
2. Real-world data and Real-world Evidence
3. Multinational Real-world Studies
4. Designing Real-World Studies
5. Propensity Score
   - Propensity Score Methods
6. Case Study I:
7. Case Study II: Optimal Time to Switch Therapy
8. Biomarker Screening in Early Phases
9. Sensitivity Analysis
10. Publishing Results

Publishing Results

- Provide estimand of interest
- Provide details of design procedure like would be described in randomize trials.
- Provide details and result of covariate balance.
- Provide results of sensitivity analysis and its meaning in the context of the results.
- Avoid prophetic extrapolation results
  - Remember that conclusions and interpretation of results are condition no unmeasured confounders

[9, 10, 11, 12, 13, 14, 4, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27]

Reference I

[1] FDA.
Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices: Draft Guidance for Industry and Food and Drug Administration Staff.
U.S. Department of Health and Human Services Food and Drug Administration Center for Devices and Radiological Health Center for Biologics Evaluation and Research, 2016.

[2] P Jepsen, S P Johnsen, M W Gillman, and H T Sørensen.
Interpretation of observational studies.
Heart, 90:956–960, 2004.

[3] Food and Drug Administration (FDA).
E 10 Choice of Control Group and Related Issues in Clinical Trials.
U.S. Department of Health and Human Services, 2011.

[4] Paul R RosenBaum and Donald B Rubin.
The central role of propensity score in observational studies for causal effects.
Biometrika, 79:688–701, 1983.

[5] Guido W. Imbens and Donald B. Rubin.
Causal Inference For Statistics, Social, and Biomedical Sciences: An Introduction.
Cambridge Press, 2015.

Reference II

[6] Paul R Rosenbaum.
Observational Studies.
Springer NY, second edition, 2002.

[7] Donald B Rubin.
Using propensity scores to help design observational studies: Application to the tobacco litigation.
Health Services and Outcomes Research Methodology, 2:169–188, 2001.

[8] Jerome Cornfield, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, and Ernst L Wynder.
Smoking and lung cancer: Recent evidence and a discusion of some questions.
Journal of National Cancer Institute, 22(1):173–203, 1959.

[9] S P Glasser, M Salas, and E Delzell.
Importance and challenges of studying marketed drugs: what is a phase iv study? common clinical research designs, registries, and self-reporting systems.
Journal of Clinical Pharmacology, 47(9):1074–1086, September 2007.

[10] Donald B Rubin.
Causal inference using potential outcomes.
, Journal of the American Statistical Association, 100(469):322–331, 2005.

Reference III

[11] J S Martitz.
On the application of probability theory to agricultural experiments: Essay on principles, section 9. in polish, but reprinted in english with discussion by t. speed and d. b. rubin.
Statistical Sccience, 5:463–480, (1923, reprinted 1990).

[12] Donald B Rubin.
Estimating causal effects of treatments in randomized and nonrandomized studies.
Journal Educational Psychology, 66:688–701, 1974.

[13] Paul R Rosenbaum and Donald B Rubin.
Constructing a control group using multivariate matched sampling methos that incorporates the propensity score.
The American Statistician, 39(1), February 1985.

[14] Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart.
Matching as nonparametric preprocessing for reducing model dependence in parametric matching as nonparametric preprocessing for reducing model dependence in parametric causal inference.
Political Analysis, 15:199–236, 2007.

[15] James M Robins, M A Hernan, and B Brumback.
Marginal structural models and causal inference inference in epidemiology.
Epidemiology, 11:550–560, 2000.

Reference IV

[16] James M Robins.
Marginal structural models.
In *1997 Proceedings of the Section on Bayesian Statistical Science*, pages 1–10, Alexandria, VA, 1998. American Statistical Association.

[17] T Sato and Y Matsuyama.
Marginal structural models as a tool for standardization.
Epidemiology, 14:680–686, 2003.

[18] Tobias Kurt, Alexander M Walker, Robrt J Glynn, K Arnold Chan, J Michael Gaziano, Klaus Berger, and James M Robins.
Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect.
American Journal of Epidemiology, 163(3):262–270, 2005.

[19] Paul R Rosenbaum and Donald B Rubin.
The bias due to incomplete matching.
Biometrics, 41:103–116, 1985.

[20] Donald B Rubin and Neal Thomas.
Affinely invariant matching methods with ellipsoidal distribution.
The Annal of Statistics, 20(2):1079–1093, 1992.

Reference V

[21] Jasjeet S. Sekhona and Richard Grieve.
A new non-parametric matching method for bias adjustment with applications to economic evaluations.
iHEA 2007 6th World Congress: Explorations in Health Economics Paper; Available at SSRN: http://ssrn.com/abstract=1138926, 2008.

[22] Jasjeet S Sekhon.
Multivariate and propensity score matching software with automated balance optimization: The matching package for r.
Journal of Statistical Software, 42(7):1–52, 2011.

[23] Alexis Diamond and Jasjeet S. Sekhon.
Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies.
Review of Economics and Statistics, 95(3):932–945, July 2013.

[24] Kao Tai Tsai and Karl Peace.
Analysis of subgroup data of clinical trials.
Journal of Causal Inference 2013;, 1(2):193–207, 2013.

Reference VI

[25] D Y Lin, B M Psaty, and R A Kronmal.
Assessing the sensivity of regression results to unmeasured confounders in observational studies.
Biometrics, 54(3):948–963, 1998.

[26] Paul R RosenBaum and D B Rubin.
Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome.
Journal of the Royal Statistical Society Series B (Methodological), 45(2):212–218, 1983.

[27] Elizabeth A. Stuart.
Matching methods for causal inference: A review and a look forward.
Statistical Science, 25(1):1–21, 2010.